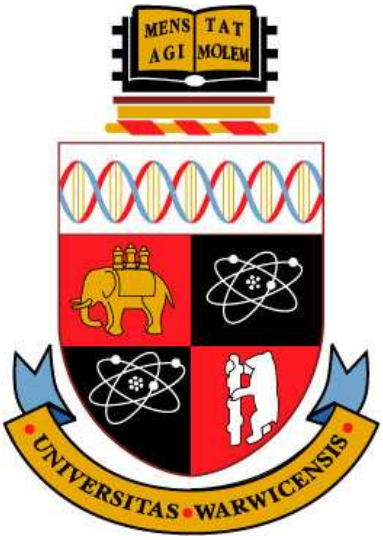

Sequential Sample Size Re-estimation in Clinical Trials with Multiple Co-primary Endpoints

by

Lupetu Ives Ntambwe

Thesis

Submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy in Health Sciences

Department of Medicine

March 2003, revised June 2014

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank everyone who has supported me through the completion of this thesis. Firstly, I would like to show gratitude to my supervisor, Prof. Nigel Stallard, for teaching me about an interesting subject and for giving me valuable support and guidance throughout my PhD studies. I have also received valuable support from the supervision of Dr. Nicholas Parsons, who has provided valuable feedback on the manuscript. I must also mention Prof. Tim Friede for supervision at the beginning of my PhD studies.

I am grateful to the Engineering and Physical Sciences Research Council (EPSRC) and Novartis for the financial support of this PhD project. I am particularly grateful to Novartis for allowing me to present my works in Basel, Switzerland, towards the end of the project.

I would also like to thank all the administrative staff of the Warwick Medical School for making sure I was comfortable.

To my parents, wife, children, brothers and sisters, thank you for the love and encouragement you have always given me. Finally, I thank God for protecting me through the period of my studies.

Declarations

I declare that the work in this thesis is all my own, except where I have stated otherwise. I also confirm that this thesis has not been submitted elsewhere for examination.

Abstract

In this thesis, we consider interim sample size adjustment in clinical trials with multiple co-primary continuous endpoints. We aim to answer two questions: First, how to adjust a sample size in clinical trial with multiple continuous co-primary endpoints using adaptive and group sequential design. Second, how to construct a test in order to control the family-wise type I error rate and maintain the power, even if the correlation ρ between endpoints is not known.

To answer the first question, we conduct K different interim tests, each for one endpoint and each at level α/K (i.e. Bonferroni adjustment). To answer the second question, either we perform a sample size re-estimation in which the results of the interim analysis are used to estimate one or more nuisance parameters, and this information is used to determine the sample size for the rest of the trial or the inverse normal combination test type approach; or we conduct a group sequential test where we monitor the information, and the information is adjusted to allow the correlation ρ to be estimated at each stage or the inverse normal combination test type approach.

We show that both methods control the family-wise type I error α and maintain the power and that the group sequential methodology seems to be more powerful, as this depends on the spending function.

Chapter 1

Introduction and Background

1.1 Introduction

The following problem is the main focus of this thesis: Suppose we have a study with two treatment groups, E (experimental) and C (control), which are to be compared in a parallel group randomised phase III clinical trial, and the same study has also K co-primary endpoints. We want to control the family-wise type I error rate, which is defined as the probability of falsely rejecting at least one null hypothesis among K hypotheses. We also want to control the power, knowing that it will depend on some nuisance parameters. Somehow we want to use interim data to modify the sample size to fix the power.

To resolve this problem, we will consider statistical methods for dealing with interim data. They will be sample size re-estimation, the group sequential approach and the inverse normal combination test procedure. Before reviewing these methods in more detail in the context of a single endpoint in the next chapter, this chapter gives some background on the concepts used throughout this thesis. Section 1.2 describes some of these concepts, Section 1.3 presents the concept of multiple endpoints, Section 1.4 describes the general framework of the analysis and Section 1.5 presents a summary.

Chapter 2 provides a literature review of the methods in the context of a single end-

point. Section 2.1 provides a background on the sample size re-estimation procedure with a single endpoint. This method enables the use of interim analyses of data to estimate one or more nuisance parameters. Following each interim analysis, this estimate is used to determine the sample size for the remainder of the trial. Section 2.2 describes the inverse normal combination test method in the context of a single endpoint. This procedure enables the combination of interim data and final data at the final analysis. As is described, this analysis method can be used along with the sample size re-estimation approach. In Section 2.3, the group sequential design method is described, again in the setting of analysis of a single endpoint. This method allows a series of interim analyses to be conducted. As described below an error spending function can be used to ensure type I error control. Finally, in Section 2.4, the group sequential design inverse normal combination test method with a single endpoint is presented. This approach integrates the inverse normal combination test method into classical group sequential testing approach.

Chapter 3 presents the idea of sample size reestimation in the context of multiple co-primary endpoints. Section 3.1 presents a sample size reestimation approach for this setting, Section 3.2 describes the inverse normal combination tests method with multiple co-primary endpoints with sample size re-estimation and Section 3.3 presents a summary of the findings.

Chapter 4 describes group sequential designs in the context of multiple endpoints. The method uses specified stopping rules and a spending function based on the information at each interim analysis. This information is adjusted to allow for the estimated correlation, ρ, between test statistics at each stage.

Chapter 5 presents the group sequential inverse normal combination tests procedure with multiple co-primary endpoints. The method integrates the concept of an inverse normal combination tests approach into the group sequential designs procedure described in

Chapter 4.

Finally, Chapter 6 provides a discussion of the main features of the new methodologies introduced, summarises their properties and offers a conclusion. Limitations of the current work and suggestions for further work are also highlighted in this chapter.

1.2 Concepts

This section gives background and introduces some concepts and statistical tools required in the rest of this thesis. Notation and key terms are defined in Subsection 1.2.1, background on clinical trials is given in Subsection 1.2.2, whilst Subsections 1.2.3 to 1.2.8 give details of the normal distribution, nuisance parameters, one-sided tests, fixed sample Z-tests, fixed sample t-tests and p-values respectively. Finally, motivation for sequential analysis is given in Subsection 1.2.9.

1.2.1 Notation and Definition of terms

The following notation is used throughout this thesis.

(i) *K* represents the total number of co-primary endpoints; there are K endpoints. These will generally be labelled as k = 1,2,...,K.

(ii) J denotes total maximum number of stages or looks. The interim stages will generally be indexed by j, j = 1,...,J.

(iii) E denotes the experimental randomised, independently, identically distributed group and C represents the control randomised, independently, identically distributed group.

(iv) n_{Ej} (n_{Cj}) denotes the total number of randomised patients in group E (C) up to and including the j^{th} stage and n_{EJ} (n_{CJ}) represents a maximum sample size specified in advance. To simplify the notation, we assume that each group has equal sample size $n_{Ej} = n_{Cj}$, and denote this by n_j. We finally assume that each group has a maximum sample size of n_J and we denote this by N.

(v) X_{ijkE} represents the random variable of the k^{th} endpoint for the i^{th} subject in group *E* at stage j ($k = 1, ..., K, i = 1, ..., n_{Ej}, j = 1, ..., J$) and X_{ijkC} represents the random

variable of the k^{th} endpoint for the i^{th} subject in group *C* at stage j ($k = 1, ..., K, i = 1, ..., n_{Cj}, j = 1, ..., J$).

(vi) E(X_{ijkE}) = θ_{kE} denotes the mean response for subjects receiving the experiment treatment E, and E(X_{ijkC}) = θ_{kC} represents the mean response for subjects receiving the control intervention C.

(vii) θ_k represents the mean difference between the two groups on the k^{th} endpoint.

(viii) $H_{0k} : \theta_k = 0$ is the null hypothesis for endpoint k.

(ix) δ_k is chosen to be a clinically important difference (or effect size) of interest and the alternative hypothesis is $H_{1k} : \theta_k = \delta_k$.

(x) σ^2_{Ek} (σ^2_{Ck}) represents the variance for response variable X_{ijkE} (X_{ijkC}). For simplicity, a common variance for endpoint k is considered i.e. σ^2_{Ek} = σ^2_{Ck} = σ^2_k.

(xi) $\rho_{k_1k_2}$ = $corr(X_{ijk_1}, X_{ijk_2})$ denotes the correlation between X_{ijk_1} and X_{ijk_2}.

(xii) Σ denotes the variance covariance matrix between endpoints.

Throughout this thesis, the term *group sequential design* will describe a sequential clinical trial that includes a series of interim analyses with the possibility of stopping the trial at each analysis. The term *sample size re-estimation* will describe a trial in which data are used at one or more interim analyses to estimate nuisance parameters with this information used to determine the sample size for the remainder of the trial. Finally, the term *inverse normal combination test* will describe a trial that uses the inverse normal p-value combination function (Bauer (1989)) to combine data from a number of stages.

1.2.2 Clinical trials

1.2.2.1 Definition and Phases of a clinical trial

Pocock (2004) defines a ***clinical trial*** as any form of planned experiment that involves patients and is designed to elucidate the most appropriate treatment of future patients with a given medical condition. In conducting a clinical trial, one uses results based on a limited sample of patients to make inferences about how a treatment should be conducted among the general population of patients who will require treatment in the future. The clinical evaluation of a new drug is usually divided into four phases. The characteristics of each phase can differ between therapeutic areas, but can roughly be outlined as follows:

- **Phase I** includes the first experiments in human beings (often healthy volunteers). Such trials are primarily concerned with drug safety, not efficacy, and the first objective is to determine how much of a drug can be given without causing serious side-effects. Studies of drug metabolism and bioavailability are also considered within this phase. Finally, in Phase I, studies of multiple doses are performed to determine the appropriate dose schedules for use in Phase II.

- In **Phase II**, the experimental treatment is first studied in patients with a view to an initial assessment of efficacy. This phase is also often used for identifying a safe and effective dose level for further development.

- In **Phase III**, a full-scale evaluation of a treatment is undertaken. The first objective is to compare the new drug, which has been shown as reasonably effective, with a placebo control or the current standard treatment(s) for the same condition in a large trial involving a substantial number of patients.

- Finally, **Phase IV**, also called the post-marketing surveillance phase, is undertaken to monitor for adverse effects and also includes additional large-scale, long-term studies

of morbidity and mortality. It is sometimes used to describe promotion exercises with the objective of bringing the new drug to the attention of a large number of clinicians.

The adaptive and sequential methods used in Phase III clinical trials are the main focus of this thesis. The primary goal in a Phase III clinical trial is usually to confirm whether or not the experimental drug (E) is efficacious compared to control (C). In a single setting, assume that the true treatment effect is θ_E for the experimental drug and θ_C for the control, and that a positive value of θ_E or θ_C indicates that the treatment has been useful to the patient. The treatment effect $\theta = \theta_E - \theta_C$ can then be assessed in a statistical hypothesis test context, where the null hypothesis $H_0 : \theta = 0$, is tested against the alternative hypothesis $H_1 : \theta > 0$. It is a regulatory requirement to control the type I error rate, *the probability of falsely rejecting the null hypothesis*, at some pre-specified level α. One-sided tests of α = 0.025 are usually required by regulators, which correspond to α = 0.05 for two-sided tests. Also, with regard to the power at a certain value of the treatment effect $\theta = \delta$, it is necessary to make sure that it is at least 1 - β. β is the type II error, i.e the *probability of failing to reject the false null hypothesis*. Popular choices for β are β = 0.1 and β = 0.2.

1.2.2.2 Clinical trials context

This thesis deals with sequential sample size re-estimation, with and without early stopping, conducted in the context of a ***randomised clinical trial***. In fact, randomised controlled trials, where patients are randomly assigned to one of several treatment groups, are the gold standard for the evaluation of a new therapy. The purpose of randomisation is to avoid bias that may arise due to unexpected variations in patients' characteristics that may result from less formal (ad hoc) allocation methods. In this type of trial, the probability of receiving a certain treatment is known, but it is not predictable what treatment each patient will receive.

The thesis also deals sample size re-estimation conducted in the context of a ***blinded randomised clinical trial***. Day and Altman (2000) explain that in controlled trials the term ***blinding***, and in particular ***double blinding***, usually refers to keeping study participants and those collecting and analysing clinical data unaware of the assigned treatment, so that they should not be influenced by that knowledge. Julious (2004) argues that blinding is important as it removes any systematic bias there may be in treatment assessment and allocation during the conduct of the trial. This forms the basis for regulatory authorities when deciding whether to approve a new drug.

1.2.3 Normal distribution

This thesis deals with sequential sample size re-estimation in clinical trials (with and without early stopping) with multiple ***continuous*** endpoints that follow normal distributions. In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution and plays a central role in statistics.

In the setting of a single random variable X, the probability density function of a normal distribution is defined by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\theta)^2}{2\sigma^2}} \tag{1.1}$$

for $\infty < x < \infty$. The probability density function is dependent on two parameters, mean θ and standard deviation σ, where $\infty < \theta < \infty$ and $\sigma > 0$.

The expected value is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \theta \tag{1.2}$$

and the variance is

$$Var(X) = \int_{-\infty}^{\infty} (x-\theta)^2 f(x) dx = \sigma^2 \tag{1.3}$$

where $f(x)$ is defined in Eq. (1.1).

Conventionally, N(θ, σ^2) is used to indicate that a random variable X follows a normal distribution with mean θ and variance σ^2.

The standard normal distribution is a special case of the normal distribution where θ = 0 and σ^2 = 1. Its density function is usually given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{1.4}$$

and its distribution function is given by

$$\Phi(x) = \int_{-\infty}^{x} \phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\mu^2}{2}} d\mu. \tag{1.5}$$

The variance defined in Eq. (1.3) is called the population variance in the sense that the concept of population can be extended to continuous random variables with infinite populations. However, in many practical situations, the true variance of a population is not known in advance and must be estimated on a sample of the population. Suppose we have a series of n measurements of a random variable X written as x_i, where i = 1, 2, ..., n. The estimate sample variance is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{1.6}$$

where $\overline{x}$ denote the sample mean of X .

The variance is calculated from the squares of the observations. This means that it is not in the same units as the observations, which limits its use as a descriptive statistic

(Bland (2000)). The answer to this is to take the square root, which will then have the same units as the observations and the mean. The square root of the variance is called the *standard deviation*, usually denoted by s. Thus

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{1.7}$$

Again, in the setting of two random variables X and Y, the probability density function of a normal distribution is defined by

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp(-\frac{1}{2}Q(x,y)) \tag{1.8}$$

where the quadratic form

$$Q(x,y) = \frac{1}{1-\rho^2}[(\frac{x-\theta_1}{\sigma_x})^2 + (\frac{y-\theta_2}{\sigma_y})^2 - 2\rho\frac{(x-\theta_1)(y-\theta_2)}{\sigma_x\sigma_y}]$$

gives the density function of a bivariate normal distribution. Note that the parameters σ_x^2, σ_y^2, and ρ must satisfy $\sigma_x^2 > 0$, $\sigma_y^2 > 0$, and $0 < \rho < 1$.

The correlation, ρ is defined as a bivariate analysis that measures the relation between two or more variables such that systematic changes in the value of one variable are accompanied by systematic changes in the other (Bobko (2001)). In statistics, the value of the correlation coefficient is +1 in the case of a perfect positive (increasing) linear relationship (correlation), - 1 in the case of a perfect decreasing (negative) linear relationship (anticorrelation). As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker (closer to uncorrelated).

Suppose we have a series of n measurements of two random variables X and Y written as x_i and y_i, where $i = 1, 2, \ldots, n$, then the sample correlation coefficient can be

used to estimate the population correlation ρ between X and Y. The sample correlation coefficient is written

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(\sum_{i=1}^{n}(x_i - \overline{x})^2)(\sum_{i=1}^{n}(y_i - \overline{y})^2)} \tag{1.9}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of X and Y.

This can also be written as

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}$$

where s_x and s_y are the sample standard deviations of X and Y as defined in Eq. (1.7).

The correlation coefficient defined in Eq. (1.9) is called Pearson correlation. It is widely used in statistics and is the one we consider in this thesis. It measure the degree of the relationship between linear related variables. Other types of correlations include Kendall rank correlation and Spearman correlation; however, they are not discussed here as they are non-parametric tests used to measure the strength of dependence between two variables for the first and the degree of association between two variables for the second (Bland (2000)).

1.2.4 Nuisance parameter

In statistics, a nuisance parameter is defined as any parameter that is not of immediate interest but must be accounted for in the analysis of those parameters that are of interest (Basu (1977)). In other words, a parameter is a nuisance parameter in the sense that we are not interested in its value, but its value modifies the distribution of our observations. For

example, if we are interested in the mean θ, the variance σ^2 defined in Eq. (1.6) and the correlation ρ_{xy} defined in Eq. (1.9) are nuisance parameters.

The variance σ^2 and the correlation ρ_{xy} may cease to be a *nuisance* if they become the object of the study. In general, a nuisance parameter is any parameter that interferes on the analysis of another.

To treat nuisance parameter in this thesis, we are going to use interim analysis to estimate the values of the nuisance parameter considered before the study begins, and this value is used to determine the parameter of interest (e.g. the sample size) for the rest of the trial or at interim stage.

1.2.5 One-sided test

Throughout this thesis, tests are conducted for the difference in the mean response of two treatments θ when observations are normally distributed with common, known (or unknown) variance σ^2.

The null hypothesis $H_0 : \theta = 0$ expresses that both treatments are equal. The alternative hypothesis $H_1 : \theta > 0$ corresponds to one treatment being greater than the other.

Suppose we have a standardised test statistic Z, which is normally distributed under H_0, and a fixed sample test rejects H_0 if $Z > c$ for a constant c. The type I error probability is defined as the probability of wrongly rejecting the null hypothesis,

$$\alpha = Pr(Z > c|\theta = 0). \tag{1.10}$$

The power of a test is the probability of rejecting the null hypothesis when it is false. It depends on the specific value of θ, denoted by δ, that is

$$Power = Pr(Z > c|\theta = \delta) = 1 - \beta \tag{1.11}$$

where δ represents a treatment difference that needs to be detected with high probability and β represents the type II error probability at $\theta = \delta$.

1.2.6 Fixed Sample Z-test

We consider a fixed sample test for a single endpoint. Let X_{iE} and X_{iC} *i* = 1,2,...,*n*, be the i^{th} observations of samples E and C. We assume that X_{iE} (X_{iC}) is normally distributed with mean θ_E (θ_C) and a common and known variance σ^2 i.e., $X_{iE} \sim N(\theta_E, \sigma^2)$ ($X_{iC} \sim N(\theta_C, \sigma^2)$), and that all observations are independent. We are interested in testing a null hypothesis that the two means are equal against an alternative hypothesis that the difference in means is a positive constant:

$$H_0 : \theta_E - \theta_C = 0$$
$$H_1 : \theta_E - \theta_C > 0$$

If *n* subjects are allocated to each treatment, the standardised statistic (see Jennison and Turnbull (2000a), p. 22) is given by :

$$\begin{aligned} Z &= \frac{1}{\sqrt{(2n\sigma^2)}}(\sum_{i=1}^{n} X_{iE} - \sum_{i=1}^{n} X_{iC}) \\ &\sim N((\theta_E - \theta_C)\sqrt{\{n/(2\sigma^2)\}}, 1). \end{aligned} \tag{1.12}$$

The information for $\theta_E - \theta_C$ is:

$$I = \frac{n}{2\sigma^2}. \tag{1.13}$$

So, under H_0 where $\theta_E = \theta_C$, $Z \sim N(0,1)$, and to satisfy the type I error probability requirement, we need:

$$Pr(Z > c|\theta = 0) = \alpha \tag{1.14}$$

where

$$c = \Phi^{-1}(1-\alpha) \tag{1.15}$$

represents the quintile of a normal distribution and Φ denotes the standard normal cumulative distribution function. The one-sided test with type I error probability α rejects H_0 if $Z > c$.

To satisfy the power requirement, we also need :

$$Pr(Z > c|\theta = \delta) = 1-\beta \tag{1.16}$$

where $Z \sim N(\theta\sqrt{I}, 1)$ and $\theta = \delta$. We denote $\mu^* = \theta\sqrt{I}$ and call it the non-centrality parameter. The power defined in Eq. (1.16) is now expressed by:

$$\begin{aligned}
1-\beta &= Pr\{Z-\mu^* > c-\mu^*|\theta=\delta\} \\
1-\beta &= Pr\{-Z+\mu^* \leq -c+\mu^*|\theta=\delta\} \\
1-\beta &= \Phi(-c+\mu^*) \\
\Phi^{-1}(1-\beta) &= -c+\mu^* \\
\mu^* &= c+\Phi^{-1}(1-\beta).
\end{aligned}$$

The sample size *n* that satisfying the power requirement can be derived by replacing μ^* by $\theta\sqrt{\frac{n}{2\sigma^2}}$ and *c* by $\Phi^{-1}(1-\alpha)$ in the above expression, giving:

$$n = \frac{2(\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta))^2\sigma^2}{\theta^2}. \tag{1.17}$$

Although we have based our test on the standardised test statistic *Z*, this is not the only possibility. Another test statistic is called the score statistic. It is defined in this case as:

$$S = Z\sqrt{I} \tag{1.18}$$

where *Z* denotes the standardised test statistic and I the information. Under H_0, the score statistic is normally distributed with mean 0 and variance *I*

$$S \sim N(0, I). \tag{1.19}$$

Under H_1, the score statistic is normally distributed with mean θI and variance *I*

$$S \sim N(\theta I, I). \tag{1.20}$$

1.2.7 Fixed sample t-test

In this subsection we consider the same problem and the same hypotheses as in Subsection 1.2.6. We assume that X_{iE} (X_{iC}) is normally distributed with mean θ_E (θ_C) and a common and unknown variance σ^2.

The common variance σ^2 can be estimated as follows;

$$S^2 = \frac{(n-1)s_E^2 + (n-1)s_C^2}{2n-2} \tag{1.21}$$

where

$$s_{E(C)}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{iE(C)} - \overline{x}_{E(C)})^2 \tag{1.22}$$

and $\overline{x}_E$ ($\overline{x}_C$) denote the sample mean of $X_{iE}(X_{iC})$.

The *t*-test is then given by

$$T = \frac{\bar{X_E} - \bar{X_C}}{S\sqrt{2/n}}. \tag{1.23}$$

Under the hypothesis of no treatment difference, T follows a *t*-distribution with *2n-2* degrees of freedom. Hence we reject the null hypothesis when $T > t_{1-\alpha,2n-2}$; where $t_{1-\alpha,2n-2}$ is define as quintile of a *t*-distribution.

Under the alternative hypothesis that there is a clinical difference $\delta > 0$, T follows a *t*-distribution with *2n-2* degrees of freedom and non-centrality parameter μ^* defined in the previous subsection as follows: $\mu^* = \frac{\delta}{\sqrt{2S^2/n}} > 0$.

The corresponding power can be written as

$$1 - \beta = P(t_{1-\alpha,2n-2} - \mu^*) \tag{1.24}$$

where P is the cumulative distribution. Practically one could use Eq. (1.17) for the initial sample size calculation and then calculate the power for this sample size using Eq. (1.23), iterating the sample size up as necessary until the required power is reached.

1.2.8 Significant tests and P-value

In Subsection 1.2.5 and Subsection 1.2.6, we explained that to carry out the test of significance, we supposed that in the population, there is no difference between the two treatments. The hypothesis of *no difference* in the population was called the *null hypothesis* H_0. We then explained that if this is not true, then the *alternative hypothesis* H_1 must be true, that there is a difference between the treatments in one direction (or the other).

The general procedure for a significant test, presented in Bland (2000), is as follows.

(i) Set up the null hypothesis and its alternative.

(ii) Find the value of the test statistic.

(iii) Refer the test to a known distribution (in our case a normal distribution) which it would follow if the null hypothesis were true.

(iv) Find the probability of a value of the test statistic arising which is as or more extreme than the one observed, if the null hypothesis were true.

(v) Conclude that the data are consistent or inconsistent with the null hypothesis.

The probability of such an extreme value of the test statistic occurring if the null hypothesis were true is often called the ***p-value***. It (p-value) is well illustrated in Figure (1.1) and is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. Its one sided form is defined mathematically as

$$p = 1 - \Phi(Z) \tag{1.25}$$

where *Z* represents the standardized test statistic defined in Eq. (1.12) and $\Phi(.)$ the cumulative standard normal distribution function defined in Eq. (1.5).

The p-value can also be computed using *t*-test defined in Eq. (1.23), that is

$$p = 1 - P(T, df) \tag{1.26}$$

where P(.) denotes cumulative distribution and df degree of freedom defined as

$$df = 2n - 2. \tag{1.27}$$

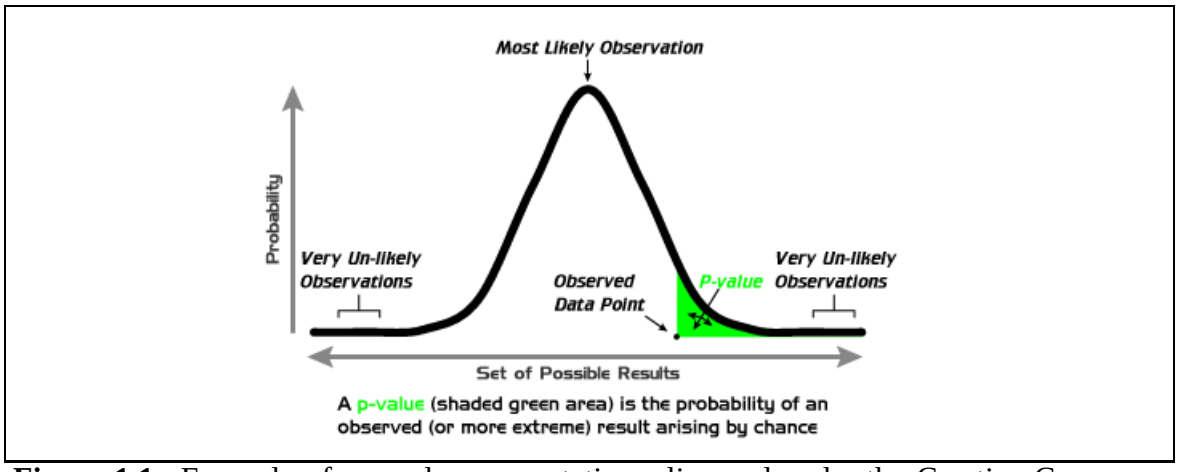

Figure 1.1: Example of a p-value computation - licensed under the Creative Commons Attribution-ShareAlike 3.0

One often rejects the null hypothesis when the *p*-value is less than the predetermined one sided significance level 0.025, indicating that the observed result would be highly unlikely under the null hypothesis (i.e., the observation is highly unlikely to be the result of random chance).

The *p*-value should not be confused with the type I error rate defined in Eq. (1.10). Even though α is also called a *significance level*, these two *significance levels* have different meanings. Their parent approaches are incompatible, and the numbers p and α cannot meaningfully be compared. The *p*-value is not the probability that the null hypothesis is true. The null hypothesis is either true or it is not; it is not random and has no probability. It is simply a measure of how likely the data is to have occurred by chance, assuming the null hypothesis is true (Bland (2000)).

1.2.9 Motivation of sequential analysis

The motivation for conducting interim assessments on accumulated data in a clinical trial has been classified as follows by several authors (e.g. Jennison and Turnbull (2000a) and Dmitrienko et al. (2005)):

Ethical. There is an ethical need to monitor the safety of patients in all treatment arms when conducting clinical trials. For example, this can be done through sequential monitoring. Less patients are exposed to an inferior treatment than for the corresponding fixed sample design when a sequential trial is stopped early for a positive effect. Patients in the trial do not have to be exposed to potential side-effects of the drugs under investigation if a group sequential trial is stopped early for futility. Likewise, the resources that would have been necessary to complete the trial can instead be used to study a different treatment in the same or another area of medical necessity.

Administrative. Some of the administrative reasons given by Jennison and Turnbull (2000a) for conducting interim analysis include the need to ensure that the experiment is conducted and executed as planned, that the subjects or experimental units are from the correct population and satisfy eligibility criteria, and that the test procedures or treatments are as prescribed in the protocol. A further administrative reason for early examination of study results is to check on assumptions made when designing the trial. For example, in an experiment where the primary response variable is quantitative, the sample size is often set assuming this variable to be normally distributed with a certain fixed variance. An early interim analysis can reveal inaccurate assumptions in time for adjustments to be made to the design. For example, Dmitrienko et al. (2005) explained that early evidence of efficacy may generate a decision to increase manufacturing spending in order to support continuing development of the experimental drug. However, to better help describe the efficacy and safety profiles of the drug, the trial may still be continued.

Economic. Sequential statistical procedures can also lead to economic benefits. For example, a trial that has a positive result may be stopped early. This is an indication that a new product can be exploited more quickly. Although, if a negative result is observed, stopping a trial early would ensure that resources are not mis-used. Jennison and Turnbull (2000a) explain that sequential procedures lead to savings in cost, time and sample

size when compared with standard fixed sample methods. The authors also added that interim analyses allow informed management choices to be made concerning the continuing allocation of limited research and development funds.

Ethics is the most important and persuasive reason for sequential clinical trials. This is why most major trials now have a data monitoring committee (DMC), whose primary responsibility is to protect the safety of patients. For example, the ICH guideline E9 Statistical Principles for Clinical Trials encourages the use of interim monitoring through group sequential methods (ICH, 1998).

Three types of data monitoring may be considered in a blinded randomisation trial (Stallard and Todd (2010)): first, administrative monitoring of clinical trial conduct and monitoring of safety data without monitoring of efficacy data; second, monitoring of efficacy data with no unblinding of treatment allocation. Stallard and Todd (2010) give an example of the estimation of nuisance parameters. Third, monitoring of efficacy data with treatment allocation unblinded to allow the estimation of the difference in efficacy between the treatments being compared. This (the third) type of monitoring presents the most ethical and statistical challenge. Stallard and Todd (2010) explain that in the second and the third type of data monitoring, administrative and safety monitoring will most likely be conducted in addition. This thesis focuses on the second and the third types of data monitoring.

When conducting sequential clinical trials, there are many aspects to consider that may differ from fixed sample trials. The books by Wald (1947), Siegmund (1985), Proschan et al. (2006), Whitehead (1997) and Jennison and Turnbull (2000a) all focus on practical and statistical aspects of interim monitoring. These books cover all the sequential design methods that will be described in this thesis and in particularly in Chapter 4.

1.3 Multiple endpoints

This section provides background on multiple endpoints in a clinical trial. It introduces various methods used to adjust for multiplicity. Subsection 1.3.1 is an introduction; Subsection 1.3.2 defines and discusses the Family-wise type I error rate; Subsection 1.3.3 describes methods for controlling family-wise type I error rate and Subsection 1.3.4 defines disjunctive power.

1.3.1 Introduction

A number of factors can influence the analysis, interpretation and conclusions drawn from a clinical trial. Among them are the disease under study, the patient population, multiple endpoints, the study design and the conduct of the study. One of the key factors that make interpretation difficult, and sometimes impossible, is the presence of multiple endpoints. Running a clinical trial with multiple endpoints may be justified by the nature of the disease and the type of questions that a clinical trial aims to investigate. For example, in patients with coronary heart disease, we may be interested in both resting and exercise ejection fractions. In blood-pressure-lowering trials, we might be interested in diastolic and systolic blood pressure or mean arterial pressure and pulse pressure. In stroke treatment, there are a number of scales used to measure recovery and no one scale is believed to assess all dimensions. In lung diseases, we may be interested in several lung function tests, such as FEV_1, FVC, PI (Pocock et al. (1987)). In behavioral studies, we may be interested in several scales for the quality of life. In patients with severe arthritis of the hip, we may be interested in the main outcome measures, hip function at 12 months after surgery, assessed using the Oxford hip score and Harris hip score (Costa et al. (2012)). These examples show that it is often inappropriate to restrict ourselves to one primary endpoint when designing or analysing a clinical trial.

Conventionally, in testing a single hypothesis, the probability of a type one error (i.e. the probability of rejecting the null hypothesis when it is true) is usually controlled at some chosen level α. For the setting considered here, this concept needs to be extended to the multiple testing situation to take account of the number of hypotheses tested.

1.3.2 Family-wise type I error rate (FWER)

1.3.2.1 Definition of family of hypotheses

Throughout this thesis, a family of hypotheses is defined as a set of hypotheses for which significance statements are considered and errors jointly controlled (Shaffer (1995)). Hochberg and Tamhane (1987) describes a family as any collection of inferences for which it is meaningful to take into account some combined measure of error.

Suppose we are considering testing a family of hypotheses, H_{0k} against a family of alternative hypotheses H_{1k}, k = 1,..,K. Suppose we do this using a series of test statistics T_k, k = 1,...,K. Suppose we also define the event that H_{0k} is rejected in preference of H_{1k} to be S_k. To adjust for multiplicity, we need to control probabilities of disjunctive events of the form $D = \bigcup_{k=1}^{K} S_k$ (Senn and Bretz (2007)). Note that the event *D* corresponds to rejecting at least one null hypothesis.

In this thesis we suppose that we are running a study for one purpose and the results are considered under one family of hypotheses. But if a study is used for different purposes, the results have to be considered under several different family configurations.

1.3.2.2 Preliminaries

Suppose K null hypotheses are of interest: H_{0k}, k =1,...,K. Suppose, as shown in Table 1.1, that U is the number of true declared null hypotheses (number of true negatives) and S denotes the number of true declared alternative hypotheses (number of true positives). V is

Table 1.1: Number of errors when testing K hypotheses

Null Hypotheses	Not Reject H_0	Reject H_0	Total
True null	U	V	K_0
True alternative	T	S	$K - K_0$
Total	W	R	K

the number of false declared null hypotheses (number of false positives or Type I error) and T denotes the number of false declared alternative hypotheses (number of false negatives or Type II error). R is the total number of null hypotheses rejections and W denotes the total number of non-rejections. K_0 is the number of true null hypotheses, an unknown parameters whereas *K* - K_0 is the number of true alternative hypotheses. U, V, S, T are not observable, whereas R and W are observable.

1.3.2.3 Individual and familywise error rate

In the context of a single hypothesis, type I error rate is the probability of rejecting the null hypothesis when it is true. It is usually controlled at some chosen level α, where α is chosen by considering the costs of rejecting a true hypothesis as compared with those of accepting a false hypothesis. It is usually set to a conventional value of 0.025 (one sided). In the context of a family of hypotheses, type I error rate is the probability of falsely rejecting at least one null hypothesis in the family or the probability of at least one error in the family. It (type I error rate) is called the *family-wise error rate* (FWER) in this setting and is defined mathematically as below:

Suppose we have K null hypotheses, denoted by $H_{01}, ..., H_{0K}$. If we use test statistics, each hypothesis may be confirmed as significant or non-significant. Table 1.1 presents a summary the test results

$$FWER = Pr(V > 0) \tag{1.28}$$

meaning the probability of making at least one type I error in the family,

or equivalently,

$$FWER = 1 - Pr(V = 0). \tag{1.29}$$

In connection to the event D (defined in Subsection 1.3.2.1), we may consider Pr(D) as the disjunctive type I error rate or FWER, that is:

$$\text{Disjunctive type I error rate} = Pr(\text{reject at least one false } H_{0k} \mid \theta_k = 0). \tag{1.30}$$

So, by assuming $FWER \leq \alpha$, the probability of making at least one type I error in the family is controlled at level α.

1.3.2.4 Control of FWER

In this thesis, we are going to test a family of hypotheses and will claim that the treatment in group E works against the one in group C if any H_{0k} is rejected, which means the FWER or disjunctive type I error rate. That is why we need to control the FWER. Some tests control the FWER only when all null hypotheses in the family are true, others control this error rate for any combination of true and false hypotheses. Hochberg and Tamhane (2001) refer to these as weak control and strong control, respectively. These concepts are discussed below.

The weak type controls the type I error only when all null hypotheses in the family are true: $H_0 = \cap_{k \in K} H_{0k}$, K_0 = K, where K_0 is the number of true null hypothesis defined in Table 1.1.

For $FWER : Pr(V > 0)$

whereas the strong type controls the type I error for any partial configuration S (defined in Table 1.1) of the null hypotheses, $K_0 \leq K$.

For FWER : $max_{S \subseteq K} P(V > 0 | \cap_{k \in S} H_{0k})$, $k = 1, \ldots, K$.

1.3.3 Methods for controlling FWER

In Subsection 1.3.2, the FWER has been defined and type of controls of the FWER have been described. In this section, we review two methods for controlling the FWER. This includes single-step and stepwise methods. In the single-step procedure, the rejection or non-rejection of a single hypothesis does not depend on the decision on any other hypothesis. Bonferroni and $\check{S}$id$\grave{a}$k methods are cited as examples. Whereas in the stepwise scenario, the rejection or non-rejection of a particular hypothesis may depend on the decision on other hypotheses. As examples, we have the Holm procedure and Hochberg method.

Although both methods are described below, only the Bonferroni method will be considered as a method for adjusting the FWER in this thesis.

1.3.3.1 Single-step procedures

Single-step procedures use the same boundary for the rejection of hypotheses. Two single-step methods are described below.

1.3.3.1.1 Bonferroni procedure

Bonferroni (1936) developed the Bonferroni procedure. It is one of the best-known and most widely used multiple-hypotheses testing procedures. It satisfactorily controls the FWER at a specified level α in the strong sense.

Let us consider a set of p-values, $p_1, ..., p_K$, to test hypotheses $H_1, ..., H_K$. The Bonferroni procedure states that if any p-value is less than α/K, $H_0 = \{H_{01}, .., H_{0K}\}$ is rejected, where H_0 is the intersection of all H_{0k}. This means that each hypothesis H_{0k} (k = 1,. . . ,K) will be individually rejected if $p_k \leq \alpha/K$, where α here is the overall level of significance. The Bonferroni inequality,

$$Pr\{\bigcup_{k=1}^{K}(p_k \leq \alpha/K)\} \leq \alpha \tag{1.31}$$

ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α (Simes (1988)).

If the K endpoints are independent,

$$\begin{aligned} Pr(\text{smallest p-value} \leq \alpha/K) &= Pr(\text{rejecting at least one } H_{0k}) \\ &= 1 - Pr(\text{not rejecting any } H_{0k}) \\ &= 1 - (1 - \alpha/K)^K \\ &= < 1 - (1 - (\alpha/K)K) \\ &= \alpha \end{aligned}$$

If the K endpoints are dependent,

$$\begin{aligned} Pr(\text{rejecting at least one } H_{0k}) &\leq \sum_{k=1}^{K} Pr(\text{reject one } H_{0k}|\theta_k = 0) \\ &= K\alpha/K \\ &= \alpha \end{aligned} \tag{1.32}$$

If alternative hypotheses are considered in which several endpoints are affected in the same

direction, Bonferroni's procedure may lack power because the rejection of the overall hypothesis is based on the smallest p-value of the K test statistics.

In practice, endpoints are usually correlated. Pocock et al. (1987) show that Bonferroni's correction practically works well for moderately correlated normally distributed endpoints with known variance and identical correlation ρ for all possible pairs within the two compared groups. The conservatism of Bonferroni's method increases as ρ increases, but Bonferroni's correction still performs well as the number of correlated endpoints increases.

1.3.3.1.2 $\check{S}id\grave{a}k$ procedure

Bonferroni's inequality was modified by Sidak (1967). Instead of testing each hypothesis at $\alpha_k = \alpha/K$, $\check{S}id\grave{a}k$ suggested using a level of significance $\alpha_k = 1-(1-\alpha)^{1/K}$. Similar to Bonferroni's approach, $\check{S}id\grave{a}k$ indicated that, for K independent endpoints:

$$\begin{aligned} Pr(\text{smallest p-value} \leq 1-(1-\alpha)^{1/K}) &= 1-\{1-[1-(1-\alpha)^{1/K}]\}^K \\ &= 1-((1-\alpha)^{1/K})^K \\ &= \alpha \end{aligned}$$

1.3.3.2 Stepwise procedures

Single step procedures use the same boundary for the rejection of hypotheses. Nevertheless, if different boundaries are assigned to different tests, testing methods may have a higher ability to preserve the FWER and to identify true alternative hypotheses. The following briefly describes three of them.

Let $H_{01}, \ldots, H_{0K}$ be a family of null hypotheses and $p_1, \ldots, p_K$ the corresponding

p-values. Consider ordering the *p*-values $p_{(1)}, \ldots, P_{(K)}$ and let the associated null hypotheses be $H_{0(1)}, \ldots, H_{0(K)}$.

1.3.3.2.1 Holm procedure

Holm (1979) proposed a method that applies in the same cases as the Bonferroni procedure but is uniformly more powerful. His step-down method proceeds as follows.

Step 1. If $p_{(1)} < \alpha/K$, reject $H_{0(k)}$ and go to Step 2; otherwise stop.

Step 2. If $p_{(2)} < \alpha/K - 1$, reject $H_{0(2)}$ and go to Step 3; otherwise stop.

...

Step K. If $p_{(K)} < \alpha$, reject $H_{0(K)}$ and stop.

The benefit of using Holm's procedure is that the tests are made more powerful (smaller adjusted p-values) while, in most cases, maintaining strong control of the FWER. The method is based on the Bonferroni inequality and is valid regardless of the joint distribution of the test statistics. However, the downside of the procedure is that the stochastic (or random) dependencies between test statistics are not taken into account.

1.3.3.2.2 Hochberg procedure

Hochberg (1988) suggested a step-up method described as follows.

Step 1. If $p_{(K)} < \alpha$, reject $H_{0(k)}$, k = 1,...,K, and stop otherwise go to Step 2.

Step 2. If $p_{(K-1)} < \alpha/2$, reject $H_{0(k)}$, k = 1,...,K - 1, and stop, otherwise go to Step 3.

...

Step K. If $p_{(1)} < \alpha/K$, reject $H0(k)$, k = 1, and stop.

The procedure is valid under independent or positively dependent p-values. Under independence, Hochberg's method is more powerful than Holm's, it maintains strong control of the FWER. However, the problems with this procedure are that the stochastic (or random) dependencies between test statistics are not taken into account and it is only valid for positively correlated test statistics.

1.3.3.2.3 Hommel procedure

Hommel (1983) suggested a method that combines both step-up and step-down procedures. The method is somewhat more powerful than Hochberg's but is more difficult to understand and carry out. The method is as follows: reject all hypotheses that have a *p*-value α/k^* where k^* is defined as

$k^* = max(k\epsilon\{1,\ldots,K\}$: $p_{(K-k+k_\star)} > \frac{k_\star\alpha}{k}$ for $k_\star = 1,\ldots,k)$.

If no maximum exists, all hypotheses are rejected (the largest p-value is then smaller than α). To illustrate this method, we consider the example given by Ekenstierna (2004) that shows that Hommel's procedure rejects more than Hochberg's:

Suppose that we have three hypotheses H_{01}, H_{02}, H_{03} and the corresponding *p*-values $p_1 = 0.012$, $p_2 = 0.015$, $p_3 = 0.0363$. Let α be 0.025. With Hommel's procedure we first calculate k^*:

For $k = 1 : p_3 = 0.0363 > \alpha = 0.025$.

For $k = 2 : p_3 = 0.0363 > \alpha = 0.025, p_2 = 0.015 > \alpha/2 = 0.0125$.

For $k = 3 : p_3 = 0.0363 > \alpha = 0.025, p_2 = 0.015 < 2\alpha/3 = 0.0167, p_1 = 0.012 > \alpha/3 = 0.000835$.

Thus in this example $k^* = max\{1, 2\} = 2$ and all hypotheses with a *p*-value $\leq \alpha/2 = 0.0125$ are rejected. The hypothesis with *p*-value $p_1 = 0.012$ is then rejected by Hommel's procedure. If Hochberg's procedure would be used instead, no hypotheses would be rejected: $p_3 = 0.0363$ is larger than $\alpha = 0.025$, $p_2 = 0.015$ is larger than $\alpha/2 = 0.0125$ and $p_1 = 0.012$ is larger than $\alpha/3 = 0.00835$.

As for Holm's and Hochberg's procedures, Hommel's method maintains strong control of the FWER. It is valid under independent or positively dependent *p*-values.

1.3.4 Disjunctive Power

Suppose we are considering testing a family of hypotheses, H_{0k}, against a family of alternative hypotheses, H_{1k}, k = 1,..,K. Suppose we do this using a series of test statistics T_k, k = 1,...,K. In connection to the event D (defined in Subsection 1.3.2.1), we may consider Pr(D) as disjunctive power, that is:

$$\text{Disjunctive power} = Pr(\text{reject at least one false } H_{0k} \mid \theta_k = \delta_k). \tag{1.33}$$

This thesis focuses mostly on the disjunctive. However, sometimes we may be interested in the probabilities of conjunctive events of the form of $C = \bigcap_{k=1}^{k} S_k$. In connection to the event C, we may consider Pr(C) as conjunctive power, that is:

$$\text{Conjunctive power} = Pr(\text{reject all false } H_{0k} \mid \theta_k = \delta_k). \tag{1.34}$$

1.4 Multiple co-primary endpoints: General framework of analysis

In this thesis, we consider methodology for situations where there are multiple continuous co-primary correlated endpoints in a clinical trial. This means that we are interested in obtaining significance in one of the endpoints. We also consider two different methods of analysing interim data: First, the sample size re-estimation method, in which data are used at one or more interim analyses to estimate nuisance parameters with this information used to determine the sample size for the remainder of the trial. Second, the group sequential designs approach, that includes a series of interim analyses with the possibility of stopping the trial at each analysis. Such a trial must be designed in advance, so as to maintain the FWER. The group sequential design method considered in this thesis is the combination of the concepts of early stopping and sample-size recalculation.

Based on concepts described previously in this chapter, the general framework for each of the two methods are defined in the following setting: There are two treatments, experimental E and control C.

1.4.1 General framework of analysis for the sample size re-estimation approach

(i) To reiterate, let X_{ikE} be the random variable of the k^{th} endpoint for the i^{th} subject in group E ($k = 1, \ldots, K, i = 1, \ldots, n_E$) and X_{ikC} be the random variable of the k^{th} endpoint for the i^{th} subject in group C ($k = 1, \ldots, K, i = 1, \ldots, n_C$).

(ii) We consider one-sided tests as described in Subsection 1.2.5.

(iii) We are interested in testing a null hypothesis that two K-dimensional mean vectors of

K co-primary endpoints are equal against an alternative hypothesis that the difference in mean vectors is a vector of positive constants:

$$H_{0k} : \theta_k = 0$$

$$H_{1k} : \theta_k = \delta_k, (\delta_k > 0)$$

where θ_k is the *k*'th element of θ (a K x 1 column vector of true means) and we are testing a family of *k* hypotheses.

(iv) Suppose T_k represents test statistics for endpoint k and c the corresponding critical value observed at the end of the trial.

(v) To reiterate, we assume that X_{ikE} (X_{ikC}) has a multivariate normal distribution leading to a multivariate normal distribution for the test statistics T_k with a mean vector θ_k and a variance vector σ_k^2.

(vi) We consider the following decision rules: if $T_k \geq c$, reject H_{0k}.

(vii) We want to control the FWER in the strong sense, that is, to have
FWER = Pr(reject any true H_{0k}) $\leq \alpha$, under any θ_k, which may combine true and false hypotheses, with at least one true hypothesis.

(viii) We want also to maintain the power, i.e Power = Pr(reject any true H_{0k}; $\theta_k = \delta_k$) = $1 - \beta$. Somehow we want to use interim data to modify the sample size to fix the power.

We generalize this framework of analysis in the following sub-section.

1.4.2 General framework of analysis for the group sequential designs method

(i) To reiterate, let X_{ijkE} be the random variable of the k^{th} endpoint for the i^{th} subject in group *E* at stage j ($k = 1, ..., K, i = 1, ..., n_{Ej}, j = 1, ..., J$) and X_{ijkC} be the random variable of the k^{th} endpoint for the i^{th} subject in group *C* at stage j ($k = 1, ..., K, i = 1, ..., n_{Cj}, j = 1, ..., J$).

(ii) We consider one-sided tests as described in Subsection 1.2.5.

(iii) At each stage, we are interested in testing a null hypothesis that two K-dimensional mean vectors of K endpoints are equal against an alternative hypothesis that the difference in mean vectors is a vector of positive constants:

$$H_{0k} : \theta_k = 0$$

$$H_{1k} : \theta_k = \delta_k, (\delta_k > 0)$$

where θ_k is the *k*'th element of θ (a K x 1 column vector of true means) and we are testing a family of *k* hypotheses.

(iv) Suppose T_{kj} represents test statistics for endpoint k at stage *j* and c_j the corresponding critical value.

(v) To reiterate, we assume that X_{ijkE} (X_{ijkC}) has a multivariate normal distribution leading to a multivariate normal distribution for the test statistics T_{kj} with a mean vector θ_k and a variance vector σ_k^2.

(vi) At each stage *j* we consider the following stopping rules: if $T_{k1} \geq c_1 or, \ldots, or T_{kj} \geq c_j$, stop at stage *j* and reject H_{0k}, otherwise continue to stage *j* + 1; where c_j represents a critical value at stage *j*.

(vii) We want to control the FWER in the strong sense, that is, to have
FWER = Pr(reject any true H_{0k}) $\leq \alpha$, under any θ_k, which may combine true and false hypotheses, with at least one true hypothesis.

(viii) We want also to maintain the power, i.e Power = Pr(reject any true H_{0k}; $\theta_k = \delta_k$) = $1 - \beta$. Somehow we want to use interim data to stop or not at stage *j* and to modify the sample size to fix the power.

1.5 Summary

This chapter provides background information on key concepts needed throughout this thesis. It gives a brief introduction in Section 1.1, describes various statistical concepts in Section 1.2, introduces the concept of multiple endpoints in Section 1.3 and Section 1.4 defines the general framework of analyses. The next chapter presents a literature review on the methods needed for this thesis.

Chapter 2

Literature review

This Chapter presents a literature review of following designs in the context of a single endpoint: sample size re-estimation, inverse normal combination test procedure and group sequential designs. Sample size re-estimation is described in Section 2.1, the inverse normal combination test in the sample size re-estimation setting in Section 2.2, group sequential designs in Section 2.3, following by inverse normal combination tests in the group sequential designs setting in Section 2.4.

2.1 Sample Size Re-estimation (SSR) with a single endpoint

This section revises characteristics of the sample size re-estimation (SSR) method. In this context, the data of the interim analyses are used to estimate one or more nuisance parameters, and this information is used to determine the sample size for the remainder of the trial. Subsection 2.1.1 introduces the section, followed by a general framework of analysis for sample size re-estimation in Subsection 2.1.2. Subsection 2.1.3 presents a formulation of the problem. Next, Subsection 2.1.4 describes unblinded methods for conducting sample size re-estimation and Subsection 2.1.5 discusses blinded methods. The final Subsection (2.1.6) describes SSR methodology for a single endpoint.

2.1.1 Introduction

The purpose of a Sample Size Re-estimation (SSR) is to obtain an *adequate sample size*. This is useful because of power testing and precision for estimation of parameters. Uncertainty during the planning stage of a clinical trial (e.g variability of a continuous endpoint) could lead to inaccurate sample size calculation due to the use of incorrect parameters. So it is necessary to use SSR method to change the sample size and to maintain power. It (SSR) is also useful for ethical, administrative and economic reasons as described in Subsection 1.2.9.

This thesis describes SSR that estimate parameters in the context of a clinical trial. In this setting, Wittes and Brittain (1990) advise that the designer of such a clinical trial should have reliable prior estimates of three classes of parameters related to the ***administration of the study*** [e.g. (i) the number of patients that the participating clinic can expect to identify; (ii) the willingness of patients and their physician to join the trial; and (iii) the recruitment rate in the clinic], the ***process of the disease*** [the variance of the outcome variable, or, for binary outcomes, the event rate in the control group; the rate of the progression of the disease in the control group during the course of the study, the rate of competing risks, etc.] and the ***effect of the treatment***. They also comment that the sample size required to detect a given effect is sensitive to all the above parameters, but their values are extremely difficult to specify accurately before the trial begins; hence they recommend performing a pilot study prior to the trial and using information obtained to adjust the parameters in order to ensure the precision of the parameters used for the design.

A sample size may be adjusted based on unblinded or blinded data from a pilot study. When performing such adjustments on unblinded data, an Independent Data Monitoring Committee (IDMC) is needed in order to preserve blindness of everybody involved in the conduct of the trial. A literature review on unblinded and blinded sample size re-estimation is given in Subsections (2.1.4) and (2.1.5) respectively, but in the next subsec-

tion we present a framework for the analysis of a SSR, followed by a description of the problem we are going to resolve in the setting of a single endpoint in Subsection 2.1.3.

2.1.2 A framework for the analysis of a SSR

In this section, we introduce an analysis framework for SSR in the context of a single endpoint. More details regarding the framework are given throughout the chapter, but this section provides a summary. Designs with sample size re-estimations are also called designs with Internal Pilot Study (IPS). This term was introduced by Wittes and Brittain (1990) to refer to the class of designs that used early observations in a trial to recalculate sample size. The sample size re-estimation method can be described as a three step procedure (Wittes and Brittain (1990)):

(i) The initial sample size calculation, leading to a provisional sample size N_0, is carried out on the basis of initial estimates of the nuisance parameters.

(ii) After recruiting $n_1 = \pi N_0$ patients (e.g., π = 0.5), the nuisance parameters are re-estimated from these observations and the sample size re-calculated to give N. This can be done using either:

- upwards adjustment, i.e. n_2 = max$(n_0, N) - n_1$ or
- unrestricted design, i.e. n_2 = max$(n_1, N) - n_1$,

and using either:

- unblinded method, i.e. a pooled estimate of the variance σ^2 or
- blinded method, i.e. use whole variance or assume difference between groups of δ.

(iii) The final analysis, including all $N = n_1 + n_2$ observations with a hypothesis test conducted using a standard t-test.

2.1.3 Formulation of the problem

Suppose that two treatment groups (E = experiment and C = control) with a normal distributed outcome are compared. Suppose also we are considering a situation of unknown and common variance σ^2. Let $H_0 : \theta_E - \theta_C = 0$ be tested against $H_1 : \theta_E - \theta_C = \delta$, $\delta > 0$. Let N_0 be the initially planned sample size per group, n_1 is the sample size per group in stage 1, n_2 is the sample size per group in stage 2 and N = $n_1 + n_2$ is the size of the entire trial per group. N_0 and n_1 are fixed numbers, while n_2 and N are chosen based on data. Also let $1 - \beta$ be a target power to detect a treatment effect $\theta = \theta_E - \theta_C = \delta$ for a given type I error α. If the variance σ^2 is known, the required sample size per group is obtained as follows

$$N_0 = \frac{2(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta))^2\sigma^2}{\theta^2}. \tag{2.1}$$

A sample size re-estimation design might be useful in the case of high uncertainty in the planning phase of a trial about the size of the variance σ^2. This can be done in an unblinded or blinded way; a number of methods have been proposed in the literature to do this, and in the following subsections we present their characteristics in more detail.

2.1.4 Unblinded methods

Initially, publications on two-stage sampling, using data of the first stage to re-estimate the sample size, were not necessarily motivated by clinical trials, therefore maintaining the treatment blind was not a primary consideration.

2.1.4.1 Stein's Method

Stein (1945) was the first to propose a two-stage procedure for determining the required sample size of the second stage.

Let n_{E1} and n_{C1}, denote the number of observations in group E and C at the stage 1; and n_E and n_C indicate the total number of observations at the end of the trial in group E and C respectively.

After n_{E1} and n_{C1} patients have been evaluated, the sample size per group is estimated based on the stage 1 estimate of variance σ^2:

$$N = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 S_1^2}{\theta^2} \tag{2.2}$$

where

$$S_1^2 = \frac{1}{2n_1 - 2}[(n_{E1} - 1)s_{E1}^2 + (n_{C1} - 1)s_{C1}^2] \tag{2.3}$$

The normal quantiles $z_{1-\alpha}$ and $z_{1-\beta}$ could be replaced by quantiles from the t-distribution with ($n_{E1} + n_{C1}$ - 2) degrees of freedom.

One continues recruiting $n_2 = N - n_1$ patients so that the required per-arm sample size N is reached and a *t*-statistic is computed at the end using the first-stage pooled variance S_1^2, defined in Eq. (2.3), at the denominator:

$$T = \frac{\overline{X}_E - \overline{X}_C}{\sqrt{2S_1^2/n}} \tag{2.4}$$

where $\overline{X}_E$ and $\overline{X}_C$ are the sample means of all *n* patients. The statistic T has a *t*-distribution with $2(n_1 - 1)$ degrees of freedom as the variance estimate uses only the first stage data.

Proschan (2009a) explained that the use of the first-stage variance is both a strength and a weakness of Stein's method. It is a strength because the denominator of Eq. (2.4) is completely determined after the first stage. Proschan argued that this would not be true if we used the usual *t*-statistic because the pooled variance at the end of the trial might differ considerably from the first stage variance. Using the first-stage variance in the denominator also guarantees that power is at least 1- β, irrespective of the true variance σ^2. It is a weakness because the first-stage estimate variance S_1^2 is less efficient than the estimate variance S^2 using all of the data.

Also Proschan (2009a) shows that the properties of Stein's procedure depend on whether S_1^2 is close to σ^2. If it is much smaller than σ^2, then the conditional type I error rate might be inflated even though the type I error rate averaged over all possible values of S_1^2 is α.

2.1.4.2 Wittes and Brittain Method (the naive t-test)

Wittes and Brittain (1990) modify Stein's procedure by using in the denominator the pooled variance estimate of all n_E + n_C observations at the end of the trial and referring this standard *t*-statistic to a *t*-distribution with n_E + n_C - 2 degrees of freedom. They call this the naive method, because it ignores the fact that the sample size was adapted.

$$S_2^2 = \frac{1}{n_E + n_C - 2}[(n_E - 1)s_{E2}^2 + (n_C - 1)s_{C2}^2] \tag{2.5}$$

and

$$T = \frac{\overline{X}_E - \overline{X}_C}{\sqrt{2S_2^2/n}} \tag{2.6}$$

Wittes and Brittain (1990) introduced the concept of the SSR in *clinical trial* setting for the two-sample *t*-test with a continuous outcome (Stein (1945) uses one-sample *t*-test).

For them, the SSR design can be described as a three-step procedure consisting of initial sample size calculation, sample size review and final analysis. The initial sample size calculation leading to a provisional sample size $\widehat{N}_0$ is carried out on the basis of initial estimates of the nuisance parameters. Friede and Kieser (2006) give an example to illustrate this by saying that earlier phases of the drug development process might inform the choice of initial estimates.

When the data of the first $n_1 = \pi N_0$ patients (e.g., π=0.5) are available, then the nuisance parameters are re-estimated from these observations, which constitute the internal pilot study. These estimates are then used for sample size recalculation with N the recalculated sample size. The sample size can then be adjusted following a predefined recalculation rule. For instance, Wittes and Brittain (1990) proposed the *restricted design* allowing only upwards adjustments of the initially planned sample size N_0, i.e. the sample size of the second stage n_2 is given by:

$$n_2 = max(N_0, N) - n_1. \tag{2.7}$$

Restricted designs require a final sample size at least as large as the originally calculated size.

Birkett and Day (1994) suggested an "*unrestricted design*" allowing downwards adjustments, i.e.

$$n_2 = max(n_1, N) - n_1. \tag{2.8}$$

Unrestricted designs permit smaller final sample sizes than originally calculated.

The final analysis includes all $N = n_1 + n_2$ observations. These observations are not analysed separately for each stage; rather all observations are pooled as in a fixed sample

size design.

To estimate S_1^2 defined in Eq. (2.3), in the unblinded fashion when n_1 patients have been recruited, the variances in the two treatment groups need to be estimated separately, then pooled, because we know that the two groups have the same variance. An Independent Data Monitoring Committee (IDMC) is needed in order to preserve blinding of all involved in the conduct of the trial.

2.1.4.3 Birkett and Day procedure

Birkett and Day (1994) extend the work of Wittes and Brittain (1990). Rather than using half of the originally planned sample size, as suggested by Wittes and Brittain (1990), Birkett and Day (1994) propose using different numbers of patients for the SSR and allowing an unrestricted design. One of their examples assumed that only the size of the internal pilot was pre-specified, not the original total sample size. They concluded that the type I error rate and power were close to target levels as long as there were at least 20 degrees of freedom to estimate the variance.

2.1.4.4 Denne and Jennison procedure

To modify the final sample size and σ^2, Denne and Jennison (1999) use a *t*-test for a two-treatment comparison based on Stein's two-stage test which involves the use of an internal pilot study. They explain that even if the estimated S^2 is less than the true variance σ^2, both Stein's procedure and Wittes and Britain's approach still use the true variance σ^2 for estimating the initial sample size N_0. To correct this, they propose a procedure that makes explicit adjustment for the random variation in the estimate of S^2. Their method controls the type I rates more closely than previously existing methods.

2.1.4.5 Wittes et al. and Coffey and Muller procedure

Wittes et al. (1999) and Coffey and Muller (1999) use a general linear model to show that the naive method is precise when restricted or unrestricted designs are considered, as it controls the type I error rate and maintains the power. They examine the impact of (i) small samples; (ii) allowing the planned sample size to decrease; (iii) the choice of internal pilot sample size; and (iv) the maximum allowable size of the second sample. Their results show that the increase in the type I error rate is often negligible, especially in restricted designs.

2.1.4.6 Kieser and Friede procedure

Kieser and Friede (2000) suggest an alternative variance estimator at the end of the trial:

$$S_*^2 = \frac{1}{n_1 + n_2 - 4}((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2). \tag{2.9}$$

where S_1^2 denotes the pooled variance before the interim analysis and S_2^2 the pooled variance after the interim analysis.

This procedure consists of replacing the pooled variance of all the data in the denominator in Eq. (2.4) and refers the test statistics to a *t*-distribution with 2(*n*-1) degrees of freedom. The procedure is called the naive method by Wittes and Brittain (1990).

2.1.4.7 Miller procedure

Miller (2005) adjusted the final unblinded variance estimate using an additive correction. The reason for this is that the sample size is determined in such a flexible way that the usual variance estimator at the end of the trial is biased. Miller derived sharp bounds for this bias.

These bounds have a quite simple form and can help in deciding if this bias is negligible for the actual study or if a correction should be done.

2.1.5 Blinded methods

In the previous subsection, sample size re-estimation methods that require unblinding of the data were discussed. In Section 4.4 of the ICH (1999) E9 guideline it states clearly that unblinding of the allocation to treatment groups for trial participants during the ongoing study is a serious concern because it could cause a bias. In this subsection, procedures that do not breach the treatment rules during the ongoing trial are reviewed. The advantage of preserving blinding is that there is no need to put in place an Independent Data Monitoring Committee (IDMC) to guarantee secrecy of interim results.

2.1.5.1 Gould and Shih procedure

The methods described above require a trial statistician to be unblinded in order to pool the separate variances from the treatment and control arms. However, the lumped variance of all observations can be computed, irrespective of treatment assignment, even without knowledge of the treatment assignments. This is why Gould and Shih (1992b) proposed variance estimators that do not require breaking the treatment blind. For them, the replacement of the combined variance in Eq. (2.2) would be the basis of sample size calculation. The logic of this approach is to take no notice of the treatment group measure and to use the estimator of the total variance as an estimator for the within-group variance. This quantity was later called the ”lumped variance” by Zucker et al. (1999):

$$S^2_{1,total} = \frac{1}{n_1 - 1} \sum_{g=1}^{2} \sum_{i=1}^{n_{1g}} (X_{1ig} - \overline{X_1})^2. \tag{2.10}$$

where g represents experiment E and control group C; and $\overline{X_1}$ is the mean of all observations in stage 1, i.e. the mean of $2n_1$ observations.

Eq. (2.10) can also be expressed as

$$S^2_{1,lumped} = \frac{1}{2n_1 - 1}\sum_{i=1}^{2n_1}(X_i - \overline{X_1})^2. \tag{2.11}$$

The idea behind the lumped (total) variance is that, at the first stage, all patients are put in one group ignoring their group of origin; then the variance is estimated as in Eq. (1.6).

Suppose that we have:

S^2_B = $\frac{SS_B}{df_B}$ denotes sample variance between group. This statistic is a measure of the variability of group means around the grand mean $\overline{X}$. It represents the mean difference or treatment difference θ where:

$SS_B = \sum_{i=1}^{g} n_i(\overline{x_{ii}} - \overline{x})^2$, denotes the sum of square between groups, where $\overline{x_{ii}}$ represents the sample mean, g describes two groups E and C and $df_B = g - 1$ denotes the degree of freedom between groups.

$S^2_W = \frac{SS_W}{df_w}$ denotes the sample variance within groups and quantifies the spread of values within groups where $SS_W = \sum_{i=1}^{g}(n_i - 1)^2$ denotes the sum of square within groups and $df_W = n - g$ denotes the degree of freedom within groups.

The total variance is the sum of within-group variance and between-group variance i.e.

$S^2_T = S^2_W + S^2_B$

Eq. (2.10) shows that the total variance is an estimate of within group variance as long as the between group variance is zero or very small. For example, in a typical clinical trial setting, the between-group variance (S^2_B) is smaller than the within-group variance

(S_W^2). This is why the total variance is often a good estimate of the within-group variance, i.e. the treatment effect θ is between 0.2 to 0.7 in a clinical trial setting.

2.1.5.2 Zucker et al. procedure

Zucker et al. (1999) developed the adjusted total variance similar to that of Wittes and Brittain (1990), see Eq. (2.12).

They let

$$S_{1,adj}^2 = S_{1,total}^2 - \frac{n_1}{4(n_1-1)}\theta^{*2}. \tag{2.12}$$

The basis of their development was to adjust the total variance for the treatment difference equal to the alternative the trial is planned for, i.e. $\theta = \theta^*$. Friede and Kieser (2001) examined the consequences of the inflation of the total variance on sample size and found this was marginal.

2.1.5.3 Gould and Shih procedure

Gould and Shih (1992b) suggested an EM algorithm-based procedure for sample size re-calculation. However, Friede and Kieser (2002) have shown that this method has some severe errors, so they recommended not to use it.

2.1.6 SSR: Methodology for a single endpoint

In this subsection, the framework of analysis described in Subsection 2.1.2 and the problem defined in Subsection 2.1.3 are used to construct a test in such a way as to control the type I

error rate and maintain the power. This will be done using a blinded method and restricted design, or upwards adjustment.

2.1.6.1 Hypotheses, test procedures and sample size calculation

We consider a two-sample situation comparing E and C. Let N_E and N_C denotes the number of randomised patients in group E and C. To simplify the notation, we assume that each group has equal sample size and denote this by N. Suppose that X_{iE} and X_{iC} ($i = 1, ..., N$) are a series of independent normal observations with means θ_E and θ_C, respectively, and common unknown variance σ^2. If the difference of means is defined as $\theta = \theta_E - \theta_C$, the corresponding test problem can be formulated as follows:

$H_0 : \theta = 0$ versus $H_1 : \theta > 0$.

The null hypotheses H_0 can be rejected at level α if $T \geq c$, where

$$T = \frac{\overline{X}_E - \overline{X}_C}{\sqrt{2S^2/N}} \tag{2.13}$$

and

$$c = t_{1-\alpha,2N-2} \tag{2.14}$$

c is the quintile of a *t*-distribution, S^2 the pooled variance estimate, $\overline{X}_E$ ($\overline{X}_C$) denote the sample mean of $X_{iE}(X_{iC})$.

The total required sample size per group for the rejection of H_0, with power $1 - \beta$ at a specified alternative $\theta = \delta$ ($\delta > 0$) is approximate by

$$N = \frac{2(\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\beta))^2\sigma^2}{\delta^2}. \tag{2.15}$$

2.1.6.2 Sample size re-estimation and test procedures

2.1.6.2.1 Sample size re-estimation

We consider the sample size re-estimation procedure described in Subsection 2.1.2. Prior to the study, a preliminary total sample size per group N_0 is calculated using an initial guess of the variance.

The first $n_1 < N_0$ observations, which constitute the internal pilot study, are used for re-estimation of the variance. The new variance estimate σ_1^2 is replaced in the sample size formula Eq. (2.15) to compute the updated total number of observations N needed to achieve the desired power. The resulting final sample size N depends on the applied sample size adjustment method as described in more details in Subsection 2.1.2. For example, if the final sample size is lower than initially planned as explained by Wittes and Brittain (1990), we have the rule N = max(N_0,N). If the final sample size is uniquely determined by the re-estimated value of the variance as suggested by Birkett and Day (1994), the rule N = max(n_1, N) applies.

After inclusion of further $n_2 = N - n_1$ patients in the second stage of the trial, the hypothesis test is performed using all N observations.

2.1.6.2.2 Type I error rate

The type I error rate is

$$Pr(T \geq c|\theta = 0) = \alpha \tag{2.16}$$

meaning that the null hypotheses H_0 can be rejected at level α if $T \geq c$, where

$$T = \frac{\overline{X}_E - \overline{X}_C}{\sqrt{2S^2/N_f}}$$

and

$$c = t_{1-\alpha,2N-2} = t\{(1-\alpha), df\} \tag{2.17}$$

c is the quintile of a *t*-distribution, df is the degree of freedom and S^2 the pooled variance estimate, $\overline{X}_E$ ($\overline{X}_C$) denote the sample mean of $X_{iE}(X_{iC})$, $i = 1, \ldots N$.

2.1.6.2.3 Power

The power of the test at $\theta = \delta$ is

$$Pr(T \geq c|\theta = \delta) = 1 - \beta \tag{2.18}$$

where N is the total required sample size for the rejection of H_0, with power $1 - \beta$ at a specified alternative $\theta = \delta$ ($\delta > 0$).

2.2 SSR Inverse Normal Combination test method with a single endpoint

Section 2.1 described sample size re-estimation methods. This section presents a method integrating the concept of the inverse normal combination test into sample size re-estimation. The method uses the inverse normal p-value combination function to combine interim data or pilot data and final data at the final analysis. After an introduction in Subsection 2.2.1, Subsection 2.2.2 discusses the two-stage combination test approach. Subsection 2.2.3 describes the inverse normal method, followed by the methodology for a single endpoint in Subsection 2.2.4.

2.2.1 Introduction

Bauer (1989) develops the *combination tests* that control the significance level even if no specific adjustment rule is pre-specified. These tests are based on the following principle: the test statistics are calculated separately from different stage data. The test decision is derived from a predefined function that combines the test statistics into a single criterion after each stage. Since the original statement of this principle, numerous multi-stage sequential procedures with interim analyses, has been proposed to allow one to modify the design of the rest of the study without compromising the overall significance level of the test decision (Brannath et al. (2002)).

This thesis will only consider the combination test method proposed by Lehmacher and Wassmer (1999). The following section describes the approach in more detail.

2.2.2 Two-stage combination test

This subsection introduces the idea of a combination test with a general combination function with a known null distribution. Suppose a null hypothesis, H_0, is tested using a two-

stage sequential design at level α against a one-sided alternative. At the design stage, we fix the design of the first phase and the test statistic to calculate a p-value, as defined in Eq. (1.25), for the test of H_0, p_1, from the sample drawn at the first stage. We also calculate p_2 from the second stage sample and consider a function C(p_1,p_2), which is used in the case where a second stage is performed to combine p_1 and p_2. We also fix in advance early decision boundaries α_1 and β_1 for $0 \leq \alpha_1 < \alpha < \beta_1 \leq 1$, with the following stopping rules.

If at the first stage:

$$p_1 \leq \alpha_1, \text{ then we reject } H_0$$
$$\text{if } p_1 > \beta_1, \text{ then we accept } H_0.$$

In both cases, we stop the trial. Furthermore,

$$\text{if } \alpha_1 < p_1 < \beta_1,$$

then we proceed to the second stage. In this case, all of the information collected at the first stage can be used to design the second stage.

At the second stage, if $C(p_1, p_2) < c$, H_0 is rejected,

c is determined by α, α_1, β_1, and the form of C(p_1, p_2) is selected before the study begins to control the α level so that

$\alpha_1 + \int_{\alpha_1}^{\beta_1} \int_0^1 1_{\{C(X,Y)\leq c\}}\, dxdy = \alpha$, where $1_{\{C(x,y)\leq c\}} = 1$ if $C(X, Y) \leq c$ and 0 otherwise.

The assumptions underlining the two-stage combination test are that the function C(p_1, p_2) is increasing in both arguments, strictly increasing in at least one argument and left continuous in p_2 (Brannath et al. (2002)). These properties must hold for all $p_1 \in]\alpha_1, \beta_1]$ and $p_2 \in [0, 1]$.

The distribution of the p_1 and p_2 under H_0 satisfies $Pr_{H_0}(p_1 \leq \alpha) \leq \alpha$ and $Pr_{H_0}(p_2 \leq \alpha | p_1) \leq \alpha$, for all $0 \leq \alpha \leq 1$.

These assumptions tell us that the distribution of p_1, and the conditional distribution of p_2 given p_1, are stochastically larger than or equal to the uniform distribution on [0,1] (Brannath et al. (2002)). This applies whenever any independent sample units are recruited at different stages and tests are applied that control the type I error probability for any significance level α chosen in advance.

Several combination functions exist in the literature, but this thesis covers only the weighted inverse normal method proposed by Mosteller and Bush (1954), and Lehmacher and Wassmer (1999). Their combination function is defined by:

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)] \tag{2.19}$$

where $0 < w_j < 1$ (j = 1,2) and:

$$w_1^2 + w_2^2 = 1. \tag{2.20}$$

The following section provides the characteristics of the two-stage combination test as proposed by Lehmacher and Wassmer (1999) in more detail.

2.2.3 Two stage Inverse Normal method

Lehmacher and Wassmer (1999) proposed an adaptive version of the group-sequential test which is sometime called the inverse normal method. Group-sequential tests are defined in more detail in the next section, however this subsection presents a method integrating the

concept of the inverse normal combination test into the sample size re-estimation methods described in Section 2.1. The framework of analysis is presented as follows.

2.2.3.1 Framework of analysis

2.2.3.1.1 Step 1

In step 1, we consider the sample size re-estimation procedure described in Subsection 2.1.2. Prior to the study, a preliminary total sample size N_0 is calculated using an initial guess of the variance.

2.2.3.1.2 Step 2

In step 2, the first $n_1 < N_0$ observations, which constitute the internal pilot study, are used for re-estimation of the variance. The new variance estimate σ_1^2 is used to calculate the *t*-test at stage 1, that is

$$T_1 = \frac{\overline{X}_{E1} - \overline{X}_{C1}}{\sqrt{2\sigma_1^2/n_1}}$$

where $\overline{X}_{E1}$ and $\overline{X}_{C1}$ denote sample mean for X_{iE1} and X_{iC1} , $i = 1, \ldots, n_1$ respectively. T_1 is then used to calculate the *p*-value at stage 1 as defined in Eq. (1.26), that is

$$p_1 = 1 - P(T_1, df_1) \tag{2.21}$$

where P(.) represents the cumulative distribution and df_1 the degrees of freedom at stage 1 as defined in Eq. (1.27)

$$df_1 = 2n_1 - 2. \tag{2.22}$$

The new variance estimate σ_1^2 is then replaced in the sample size formula Eq. (2.15) to compute the updated total number of observations N. The resulting final sample size N depends on the sample size adjustment method as described in more details in Subsection 2.1.2.

Furthermore, in step 2, an additional $n_2 = N - n_1$ observations are used for re-estimation of the variance. The new variance estimate σ_2^2 is used to calculate the *t*-test at stage 2, that is

$$T_2 = \frac{\overline{X}_{E2} - \overline{X}_{C2}}{\sqrt{2\sigma_2^2/n_2}}$$

T_2 is then used to calculate the p-value at stage 2, that is

$$p_2 = 1 - P(T_2, df_2) \tag{2.23}$$

where P(.) represents the cumulative distribution and df_2 is defined as

$$df_2 = 2n_2 - 2. \tag{2.24}$$

2.2.3.1.3 Step 3

In step 3, which constitutes the final analysis, the evidence from stage 1 and stage 2 are combined via the weighted inverse normal functions of the observed p_1 and p_2. The resulting test statistic

$$B = w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2) \tag{2.25}$$

is used to perform a hypothesis test, where w_1 and w_2 represent the weights chosen independently of the observed data. More details on when the null hypothesis H_0 is rejected

based on the value of the test statistic *B* are addressed below.

2.2.3.2 Characteristics of the Inverse normal combination test

Under the null hypothesis, p_j, j = 1,2 in Eq. (2.25) is uniformly distributed on (0, 1), that is

$$p_j \sim U[0,1]. \tag{2.26}$$

If this is so, then

$$\Phi^{-1}(1-p_j) \tag{2.27}$$

follows the standard normal distribution, and for any constant pre-defined weight w_j, $w_j\Phi^{-1}(1-p_j)$ is normally distributed with mean 0 and variance w_j^2, that is

$$w_j\Phi^{-1}(1-p_j) \sim N(0, w_j^2), j = 1, 2. \tag{2.28}$$

If the weights, w_j, j = 1,2, are determined in advance, the $w_j\Phi^{-1}(1-p_j)$ term are independent, so under the null hypothesis, $B = w_1\Phi^{-1}(1-p_1)+w_2\Phi^{-1}(1-p_2)$ is normally distributed with mean 0 and variance $(w_1)^2 + (w_2)^2$, that is,

$$B \sim N(0, (w_1)^2 + (w_2)^2). \tag{2.29}$$

An other way of presenting the test statistics proposed by Lehmacher and Wassmer (1999) is:

$$B = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \Phi^{-1}(1-p_j), J = 2 \tag{2.30}$$

where $\frac{1}{\sqrt{J}}$ represents the weight w_j, j = 1,2 as in Eq. (2.25).

Whitehead (2010) explains that the choice of weights w_j has to be made in advance. These can all be set to 1 such as $(w_1)^2 + (w_2)^2 = 1$, or they can be chosen to reflect the information contained in each new batch of data. The author explains that the difficulty in implementing the later strategy is that many adaptive designs are used precisely to allow flexibility in the choice of sample size for each stage of the trial, and so the amount of information to be collected will not be known in advance.

If the weights are set to 1 i.e, $\sum_{j=1}^{2} w_j^2 = 1$, as suggested by Whitehead (2010), the test statistics *B* defined in Eq. (2.25) and *Z* defined in Eq. (1.12) have the same distribution under the null hypothesis i.e. N(0,1), consequently *B* can use the same critical value as *Z* at the final analysis, that is

$$c = \Phi^{-1}(1 - \alpha).$$

In the following section, we present methodology of this method.

2.2.4 SSR Inverse Normal Combination test method: Methodology for a single endpoint

In this subsection, the framework of analysis described in Subsection 2.2.3.1 is used to construct a test in such a way as to control the type I error rate. However, power considerations are covered here as the specification and interpretation of alternative hypotheses is more difficult to define in general (Whitehead (2010)).

2.2.4.1 Hypotheses and Test procedures

We consider the same problem as in Subsection 2.1.6.1. Suppose a two-sample situation comparing E and C. Let N_E and N_C denotes the number of randomised patients in group E and C. To simplify the notation, we assume that each group has equal sample size and denote this by N. Suppose that X_{iE} and X_{iC} ($i = 1, ..., N$) are a series of independent normal observations with means θ_E and θ_C, respectively, and common unknown variance σ^2. If the difference of means is defined as $\theta = \theta_E - \theta_C$, the corresponding test problem can be formulated as follows:

$H_0 : \theta = 0$ versus $H_1 : \theta > 0$.

In Subsection 2.2.3.1, we have shown that at the final analysis, the test statistic is obtained by combining the observations at stage 1 and the new observations at stage 2 via the weighted inverse normal functions of the observed p-values at stage 1 and 2, that is

$$B = w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)$$

where w_1 and w_2 are chosen independently of the observed data. In the same subsection, we also show that if the weights, w_j, j = 1,2, are determined in advance, the terms $w_1\Phi^{-1}(1 - p_2)$ and $w_2\Phi^{-1}(1 - p_2)$ are independent, so under the null hypothesis,

$$B \sim N(0, (w_1)^2 + (w_2)^2) = N(0, 1).$$

The null hypotheses H_0 can be rejected at level α if $B \geq c$, where *c* represents the quintile of a normal distribution as defined in Eq. (1.15).

A type I error probability is:

$$Pr(B \geq c|\theta = 0) = \alpha. \tag{2.31}$$

So to control the type I error rate, one must use the critical value *c* to satisfy Eq. (2.31).

The specification and interpretation of alternative hypotheses is more difficult to define in general, although in special cases it will be possible to make some form of power requirement (Whitehead (2010)). However, in this thesis, we consider using the sample size of the SSR method and the inverse normal combination test statistics to maintain the power. This will be checked by simulations in the setting of multiple co-primary endpoints.

2.2.4.2 SSR Inverse Normal Combination test: Motivation

The main reason for choosing inverse normal combination test method in SSR setting is that analytically, it has been proven that the combination test method maintains the type I error rate for any possibly data-driven choice of sample sizes. To illustrate this, consider first a fixed sample size setting with an interim look halfway through as suggested by Proschan (2009b).

If the data are independent and identically distributed (iid) normal with known variance σ^2, the most appropriate way to combine the two independent halves is to combine the Z-scores Z_{stage1} and Z_{stage2} using $Z = \frac{1}{\sqrt{2}}(Z_{stage1}) + \frac{1}{\sqrt{2}}(Z_{stage2})$, which is the usual Z-statistic on the full sample. A more general method that works for other types of data with or without the assumption of known variance is to combine *p*-values using the inverse normal method proposed by Lehmacher and Wassmer (1999): $B = w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)$.

Now suppose that after looking at results from the first half, we decide to change the second stage sample size. How does this affect the joint distribution of the *Z*-scores or *p*-values? The following result, proposed by Proschan (2009b), underlying adaptive two-stage procedures shows that it does not.

Result 1: If p_1 and p_2 are uniformly distributed on (0,1) in a fixed sample setting with iid data, the same is true if the second stage sample size depends on first stage data. Therefore, any α-level rejection region in a fixed sample setting remains level α in the adaptive sample size setting.

Proschan (2009b) gives some immediate applications of *Result 1* as follows.

(i) *Result 1* shows that p_1 and p_2 are uniformly distributed on (0,1) and B is normal distributed with mean 0 and variance 1 even when the second stage sample size depends on stage 1 data.

(ii) *Result 1* shows that the first and second stage *Z*-scores remain iid N(0,1) under H_0 even though the second stage sample size depends on results from the first stage i.e, $Z = \frac{1}{\sqrt{2}}(Z_{stage1}) + \frac{1}{\sqrt{2}}(Z_{stage2})$ has a standard normal distribution, so regardless of whether the second stage sample size is changed, we may refer Z to a standard normal distribution at the end.

2.3 Group Sequential Designs with a single endpoint

In Section 2.1, the SSR method was described as one type of clinical trial design in which the data of the interim analyses are used to estimate one or more nuisance parameters, and this information is used to determine the sample size for the remainder of the trial. Another type of clinical trial design is the group sequential design (GSD). In this context, the accumulating data are analysed at a series of interim analyses. They allow a trial to be stopped or continued at interim analyses. Further details of the methods described here are given by Wald (1947), Siegmund (1985), Whitehead (1997), Proschan et al. (2006) and Jennison and Turnbull (2000a). After an introduction in Subsection 2.3.1, Subsection 2.3.2 describes elements of sequential methodology. Subsection 2.3.3 presents stopping boundary calculations, followed by a brief description of post trial analysis in Section 2.3.4.

2.3.1 Introduction

A sequential clinical trial is one in which accumulating data are analyzed at a series of interim analyses. One potential type of a sequential trial is called a Group Sequential Design (GSD). It allows a trial to be stopped at an interim time point during a planned sequence of analyses. In this setting, the trial might be stopped with the conclusion that the experimental treatment is effective, to abandon the trial, or otherwise be continued to the next interim analysis (Jennison and Turnbull (2000a)). Such trials must be designed in advance, with the specified design adhered to, so as to maintain the overall type I error rate.

Sequential monitoring on accumulated data was initially focused on applications in quality control, which gained importance during World War II when it was essential to make sure that ammunition was of appropriate quality. Wald (1945) and Wald (1947) proposed the sequential probability ratio test (SPRT) for testing the simple null hypothesis $H_0 : \theta = \theta_0$

against the alternative $H_1 : \theta = \theta_1$. In Wald's SPRT, the only decision to be made is whether to terminate or continue the trial. This classical sequential design is called an *open* plan because there is no maximum sample size. Wald and Wolfowitz (1948) showed that under certain assumptions, the SPRT is optimal in the sense that it minimises the expected sample size. However, the SPRT leads to an issue, which is that there is no maximal sample size at which sampling is guaranteed to stop. Consequently, the distribution of the sample size can be quite skewed with a large variance. Armitage (1957) introduced the *closed* sequential design to impose a limit on the sample size. Later, McPherson and Armitage (1971) proposed a theory of repeated significance tests on accumulating data that is similar to the *closed* sequential design. Despite the savings in sample size, the need for constant data monitoring and rapid response measures was not interesting. Group sequential designs were later developed to avoid some of the problems of classical sequential designs.

2.3.2 Elements of a sequential method

Whitehead (1999) proposes four fundamental elements of any sequential method, allowing the significance of the treatment difference to be evaluated and its magnitude to be estimated. In the case of a single endpoint, two treatments, and considering a frequentist approach of analysis, the key elements are as follows:

(i) A parameter θ, which is an unknown population characteristic and expresses the benefit of E over C in terms of efficacy.

(ii) A statistic that provides information on the size of this benefit based on the sample of data available at an interim analysis, and a second statistic that gives the amount of information about θ contained in the sample.

(iii) A stopping rule, which determines, on the basis of the observed test statistics, values at the interim analysis whether to continue or to stop a trial.

(iv) A final analysis procedure, valid for the stopping rule used, which enables one to conclude whether or not E is superior to C and provides a p-value and point estimate and confidence interval for the treatment difference at the end of the trial.

Diverse suggestions have been provided by various authors for each of these four elements. In the following subsection they are presented in more detail. The specification of the treatment difference θ and test statistics is required when either fixed sample or sequential methods are considered. A solution to the problem of the control of the type I error rate is the main reason for the specification of the third element, which is the stopping boundary, and several approaches are seen in existing literature on how to calculate it. Some of these methods are presented below. With regards to the fourth element, Stallard and Todd (2010) clarify that the final analysis is very important in the interpretation of the results from the sequential trial. Statistical methodology in this area has generally fallen behind that for the construction of stopping rules. In practice, the use of an appropriate final analysis has often been neglected, in part due to the lack of availability of suitable software. Methods for the validity of analysis have been developed, and the analysis can now be conducted using commercially available software for some settings, but not all.

2.3.2.1 Parametrisation of treatment difference

When designing a sequential clinical trial, one must first select an appropriate primary outcome measure of the treatment efficacy. This is because the importance of any clinical trial will be significantly enhanced if a single primary analysis is specified in the protocol, and is later found to show significant benefit from the experimental treatment (Whitehead (1999)). This is true even in the case of a trial with a fixed sample size. This measure should be chosen on the basis of clinical relevance, ease and accuracy of measurement, and familiarity to clinicians, as explained by Stallard and Todd (2010).

After defining the primary endpoint, one must then choose an associated *parameter* that measures the difference between the experimental and control treatments. This will be influenced by the type of data collected on the primary endpoint, the ability to interpret the parameter (eg. ratio or difference) and the accuracy of the resulting analysis. In this thesis, only parallel group studies are considered. The parameter θ represents a measure of the difference between the experimental and the control treatment groups. A value of zero for θ corresponds to equivalence of the treatments, positive values correspond to an advantage for the experimental treatment and the negative values correspond to an advantage for the control.

In this thesis, only continuous outcomes are considered; for example, blood pressure. It (blood pressure outcome) represents a difference in true unknown mean blood pressure between the two groups of interest. Other classes of response exist but they are not considered here. For example, if the primary response is the time to some event, such as time from HIV infection to AIDS, the difference between treatment groups might be measured by the ratio of the hazards in the two groups or the logarithm of this ratio. If the primary response is a dichotomous variables such as success or failure, the chance of success in each group can be measured by the odds of success, and the difference between the treatment groups can be measured by the ratio of the odds in the two groups.

2.3.2.2 Test statistics and distribution theory

Suppose that we are interested in repeated looks at the accumulating data on the primary endpoint with repeated hypothesis testing. At each interim analysis we will base inference on some calculated test statistics. We need to think about the distribution for the data and the test statistics. In the following, an example of a single normal sample with known variance is given followed by a more general case of two normal samples with nuisance

parameters. Throughout this chapter, the notation introduced in Section 2.1 is still valid, but the index k is omitted because we are in the scenario of a single endpoint.

2.3.2.2.1 Single normal sample with known variance

Assume we have a sequence of independent identically distributed observations $X_i \sim N(\theta, 1)$, ($i = 1, ..., n_j$, j = 1,..,J) and we want to draw inference regarding θ and test $H_0 : \theta = 0$. At look *j*, we have $x_1...x_{n_j}$ observations (including those from previous looks so, that $n_j \geq n_{j-1}$).

The likelihood for θ at look *j* is:

$$\begin{aligned} L(\theta; x_1...x_{nj}) &= \prod_{i=1}^{n_j} L_{ij}(\theta) \\ &= \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x_i - \theta)^2\} \\ &= (\frac{1}{\sqrt{(2\pi)}})^{n_j} exp\{\sum_{i=1}^{n_j}(-x_i^2/2 + \theta x_i - \theta^2/2)\} \end{aligned} \tag{2.32}$$

where θ is the population parameter and $L_{ij}(\theta)$ is the probability or probability density of x_i. Using the likelihood function, two statistics can be derived that are useful for inference: the likelihood estimator and the test statistics.
Hence, the log-likelihood is:

$$\begin{aligned} l(\theta) &= -n_j log(2\pi)/2 - \sum_{i=1}^{n_j} x_i^2/2 + \theta \sum_{i=1}^{n_j} x_i - \theta^2 n_j/2 \\ &= constant + (\sum_{i=1}^{n_j} x_i)\theta - (\frac{n_j}{2})\theta \end{aligned} \tag{2.33}$$

where $\sum_{i=1}^{n_j} x_i$ represents a sufficient statistic for θ. Also, $l(\theta)$ is a quadratic with linear and quadratic coefficients $\sum_{i=1}^{n_j} x_i, -n_j/2$.

Assume $S_j = \sum_{i=1}^{n_j} x_i$. S_j will be the basis of our inference, i.e S_j is normally distributed with mean θn_j and variance n_j.

$$S_j \sim N(\theta n_j, n_j)$$

if $j_1 \leq j_2$, cov(S_{j_1}, S_{j_2}) = n_{j_1}.

Thus S_1,...S_J have a joint multivariate normal distribution:

$$\begin{pmatrix} S_1 \\ . \\ . \\ . \\ S_J \end{pmatrix} \sim N \left(\begin{pmatrix} \theta n_1 \\ . \\ . \\ . \\ \theta n_J \end{pmatrix} , \begin{pmatrix} n_1 & n_1 & . & . & . & n_1 \\ n_1 & n_2 & & & & n_2 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ n_1 & n_2 & . & . & . & n_J \end{pmatrix} \right) \tag{2.34}$$

The standardised test statistic at stage *j* is:

$$Z_j = S_j / \sqrt{var(S_j)} = S_j / \sqrt{n_j}$$

$$Z_j \sim N(\theta \sqrt{n_j}, 1)$$

Under H_0, Z_j is normal distributed with mean 0 and variance 1, i.e. $Z_j \sim N(0, 1)$

$cov(Z_{k_1}, Z_{k_2})$ = $cov(\frac{S_{j_1}}{\sqrt{n_{j_1}}}, \frac{S_{j_2}}{\sqrt{n_{j_2}}})$ = $\frac{n_{j_1}}{\sqrt{n_{j_1} n_{j_2}}}$ = $\sqrt{\frac{n_{j_1}}{n_{j_2}}}$, $j_1 \leq j_2$

$Z_1, ..., Z_J$ have a joint multivariate normal distribution:

$$\begin{pmatrix} Z_1 \\ . \\ . \\ . \\ Z_J \end{pmatrix} \sim N\left(\begin{pmatrix} \theta\sqrt{n_1} \\ . \\ . \\ . \\ \theta\sqrt{n_J} \end{pmatrix} , \begin{pmatrix} 1 & \sqrt{\frac{n_1}{n_2}} & . & . & . & \sqrt{\frac{n_1}{n_J}} \\ \sqrt{\frac{n_1}{n_2}} & 1 & & & & \sqrt{\frac{n_2}{n_J}} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \sqrt{\frac{n_1}{n_J}} & \sqrt{\frac{n_2}{n_J}} & . & . & . & 1 \end{pmatrix} \right) \tag{2.35}$$

An alternative test statistic at stage j is:

$\widehat{\theta_j} = S_j/var(S_j) = S_j/\sqrt{n_j}$

$\widehat{\theta_j} \sim N(\theta, 1/\sqrt{n_j})$

$cov(\widehat{\theta_{j_1}}, \widehat{\theta_{j_2}}) = cov(\frac{S_{j_1}}{n_{k_1}}, \frac{S_{j_2}}{n_{j_2}}) = \frac{n_{j_1}}{n_{j_1} n_{j_2}} = \frac{1}{n_{j_2}}$, $j_1 \leq j_2$

$\widehat{\theta_1}, ..., \widehat{\theta_J}$ have a joint multivariate normal distribution:

$$\begin{pmatrix} \widehat{\theta_1} \\ . \\ . \\ . \\ \widehat{\theta_J} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ . \\ . \\ . \\ \theta \end{pmatrix} , \begin{pmatrix} \frac{1}{n_1} & \frac{1}{n_2} & . & . & . & \frac{1}{n_J} \\ \frac{1}{n_2} & \frac{1}{n_2} & & & & \frac{1}{n_J} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \frac{1}{n_J} & \frac{1}{n_J} & . & . & . & \frac{1}{n_J} \end{pmatrix} \right) \tag{2.36}$$

The above results show that if we want to test H_0, test statistics S_j, Z_j, $\widehat{\theta_j}$ can effectively be used interchangeably using the appropriate distribution.

Note that in the Eq. (2.34), defining the distribution of the S_j statistics,
$S_1 \sim N(\theta n_1, n_1)$

$S_j - S_{j-1} \sim N((n_j - n_{j-1})\theta, (n_j - n_{j-1}))$, independent of S_{j-1}

That is, the increments S_1, $S_2 - S_1$, $S_j - S_{j-1}$ are independently distributed.

2.3.2.2.2 A more general distribution

The results from the previous subsection can be made more general. For instance, for non-normal data, unknown nuisance parameters or the comparison of two samples. If we wish to draw inference regarding θ and test $H_0 : \theta = 0$, we need to specify a model in terms of the parameter of interest θ, and obtain the (profile) log-likelihood for θ.

Consider a Taylor series expansion for $l(\theta)$ at $\theta = 0$:

$$l(\theta) = l(0) + \frac{dl}{d\theta} \mid_{\theta=0} \theta - \frac{1}{2}\frac{d^2l}{d\theta^2} \mid_{\theta=0} \theta^2 + ...$$

$\frac{dl}{d\theta} \mid_{\theta=0}$ is the efficient score statistic, i.e. the first derivative of the log-likelihood

- $\frac{d^2l}{d\theta^2} \mid_{\theta=0}$ is the Fisher's information statistic, i.e. minus the second derivative of the log-likelihood.

Denoting these by S and I, we can write

$$l(\theta) \approx l(0) + S\theta - \frac{I}{2}\theta^2 \text{ , where } I = -\frac{d^2l}{d\theta^2} \mid_{\theta=0}$$

This is the same as the single-sample normal log-likelihood, with S replacing the sample sum at stage *j* and I replacing n_j.

If we denote the score statistic at look *j* by S_j and use this as our test statistic, we have

$$\begin{pmatrix} S_1 \\ . \\ . \\ . \\ S_J \end{pmatrix} \sim N \left(\begin{pmatrix} \theta I_1 \\ . \\ . \\ . \\ \theta I_K \end{pmatrix} , \begin{pmatrix} I_1 & I_1 & . & . & . & I_1 \\ I_1 & I_2 & & & & I_2 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ I_1 & I_2 & . & . & . & I_K \end{pmatrix} \right) \tag{2.37}$$

at least for large samples and small θ. Jennison and Turnbull (2000a) describe this as the canonical form, and almost all group-sequential methods are based on this assumption. S_j and I_j are the same as defined in Eq. (1.18) and Eq. (1.13), respectively.
$S_1, ..., S_J$ are like points on a continuous Brownian motion process. Wald (1947) and Siegmund (1985) described other methods based on this model. Jennison and Turnbull (2000a) provided a comprehensive account of how to construct group sequential tests for a wide range of response distributions.

The standardised test statistic version for Eq. (2.37) is given by

$$\begin{pmatrix} Z_1 \\ . \\ . \\ . \\ Z_J \end{pmatrix} \sim N \left(\begin{pmatrix} \theta\sqrt{I_1} \\ . \\ . \\ . \\ \theta\sqrt{I_J} \end{pmatrix} , \begin{pmatrix} 1 & \sqrt{\frac{I_1}{I_2}} & . & . & . & \sqrt{\frac{I_1}{I_J}} \\ \sqrt{\frac{I_1}{I_2}} & 1 & & & & \sqrt{\frac{I_2}{I_J}} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ \sqrt{\frac{I_1}{I_J}} & \sqrt{\frac{I_2}{I_J}} & . & . & . & 1 \end{pmatrix} \right) \tag{2.38}$$

2.3.2.3 Stopping rules

The distribution of the test statistics introduced in Eq. (2.38) are used at each interim stage of a GSD and have this form of the stopping rule:

After stage *j* = 1, ..., J-1		
if $Z_j \geq c_j$	stop, reject H_0	
otherwise	continue to stage *j* +1	(2.39)
after stage J		
if $Z_J \geq c_J$	stop, reject H_0	
otherwise	stop, accept H_0	

where c_j represents a critical value or boundary for Z_j. Subsection 2.3.3 provides details on how to compute c_j numerically.

A type I error probability is:

$$\sum Pr(Z_1 < c_1, ..., Z_{j-1} < c_{j-1}, Z_j \geq c_j | \theta = 0) = \alpha \tag{2.40}$$

for some *j* = 1,...,J, estimated when $\{Z_1, ..., Z_j\}$ follow the null distribution of Eq. (2.38).

The power of the test at $\theta > 0$ is

$$\sum Pr(Z_1 < c_1, ..., Z_{j-1} < c_{j-1}, Z_j \geq c_j | \theta = \delta) = 1 - \beta \tag{2.41}$$

for specified $\delta > 0$, estimated when $\{Z_1, ..., Z_j\}$ follow the distribution in Eq. (2.38). The probability in Eq. (2.41) results entirely from outcomes that terminate at an analysis j with $Z_j \geq c_j$ when $\theta > 0$. For given J, α, β and $\{c_1, ..., c_J\}$, the maximum sample size can be found that satisfies Eq. (2.41) when $\{Z_1, ..., Z_j\}$ follow the distribution in Eq. (2.38).

A variety of tests (e.g. Pocock, O'Brien $\&$ Fleming, etc...) use different sequences of boundaries $\{c_1, ..., c_j\}$, but all are chosen to ensure the type I error probability is equal

to a specified value α, or the power is equivalent to a specific value $1-\beta$, when Eq. (2.38) holds. In the following section we review some of these.

2.3.2.3.1 Pocock's Test

Pocock (1977) proposes group sequential tests, where $n_{Ej} = n_{Cj} = n_j$, *j* = 1,...,J. The test satisfies Eq. (2.39) with $c_j = C_p(J, \alpha)$, where $C_p(J, \alpha)$ represents Pocock's boundary. It is constant throughout interim stages and is computed numerically using the joint distribution of the sequence of statistics $Z_1, ..., Z_j$; see Subsection 2.3.3 below for more details. The boundary $C_p(J, \alpha)$ is chosen to give overall type I error α, i.e.,

$$Pr_{(\theta=0)}(\text{Reject } H_0 \text{ at analysis j = 1, j = 2, ..., or j = J}) = \alpha \tag{2.42}$$

The power is then given by

$$Pr_{(\theta=\delta)}(\text{Reject } H_0 \text{ at analysis j = 1, j = 2, ..., or j = J}) = 1-\beta \tag{2.43}$$

The maximum sample size of the group sequential design depends on J, α, β, and is proportional to $\frac{\sigma^2}{\theta^2}$.

2.3.2.3.2 O'Brien and Fleming's Test

O'Brien and Fleming (1979) proposed a sequential method that boundary values decrease over the stages on the standardised normal *Z* scale, as defined in Eq. (1.12), to make the early stop less likely. The procedure has conservative stopping boundary values at very early stages, and boundary values at the final stage are close to the fixed-sample design.

Formally, the test follows Eq. (2.39) with $c_j = C_B(J, \alpha)\sqrt{(J/j)}$.

The power is given by:

$$P_{(\theta=\delta)}(\text{Reject } H_0 \text{ at analysis j = 1, j = 2, ..., or j = J}) = 1 - \beta \tag{2.44}$$

The maximum sample size that the group sequential design may need depends on J, α, β, and is proportional to $\frac{\sigma^2}{\theta^2}$.

In general, O'Brien $\&$ Fleming's test requires a smaller group size than Pocock's test to satisfy the same power requirement, but is less likely to stop early.

2.3.2.3.3 Spending function approach

Pocock, and O'Brien and Fleming's approaches ensure that the overall type I error is equal to the pre-specified value if test statistics have the canonical form. But all these designs require a fixed number and spacing of interim looks specified at the design stage. An alternative approach was introduced by Lan and DeMets (1983). It is based on an error spending approach and allows the data monitoring committee to change the timing and frequency of interim evaluations. Within the error spending context, interim monitoring enables the logical basis of the design thinking while allowing significant flexibility (Jennison and Turnbull (2000a)).

The tests defined in the previous section assume information sequences to be fixed. If the observed information sequence is different from the one used to derive the critical values prior to the start of the trial, the type I error will no longer be equal to α. Lan and DeMets (1983) proposed error spending designs as a way of dealing with random information sequences. In error spending designs, the cumulative type I error, partitioned into probabilities $\pi_1, ..., \pi_j$ (j = 1,...,J) is specified as a function of the observed information.

The boundary at a decision time is determined by π_j (j = 1,...,J), and by past and current decision times, but does not depend on the future decision times or the total number of decision times.

As noted by Jennison and Turnbull (2000a), if the information levels $I_1, ..., I_j$ are observed, the critical value c_j, for the standardised statistics Z_j, conditionally on $I_1, ..., I_j$ are calculated such that:

$$Pr_{(\theta=0)}(Z_1 < c_1, ..., Z_{j-1} < c_{j-1}, Z_j \geq c_j) = \pi_j \tag{2.45}$$

where π_j represents the probability of stopping at stage *j* to reject H_0 when this hypothesis is true. It is also called the error spent at stage *j*.

In the context of a maximum information trial, Jennison and Turnbull (2000a) explain that the type I error rate can be divided according to an error spending function, *f(t)*, which is non-decreasing and satisfies *f(0)* = 0 and *f(t)* = α for $t \geq 1$. The quantity *t* represents the information time. It is supposed that the maximum information I_{max} will be reached if the trial does not stop with an early decision. I_{max} and *f(t)* must be selected before the study begins. So the type I error probabilities for each stage are

$$\begin{aligned} \pi_1 &= f(I_1/I_{max}) \\ \pi_j &= f(I_j/I_{max}) - f(I_{j-1}/I_{max}). \end{aligned} \tag{2.46}$$

The critical values c_j are computed to satisfy Eq. (2.45). They do not depend on unobserved information I_{j+1},I_{j+2},...,I_J. The stopping rule defined in Eq. (2.39) is still valid here, but the only difference is the critical values c_j depend on the observed values I_1, I_2,....I_j.

Flexible spending function families exists which can also approximate group sequential boundaries. Two of them have been proposed by Lan and DeMets (1983):
The Pocock type is defined as

$$f(t_j) = \alpha log\{1 + (e-1)t_j\}, \tag{2.47}$$

O'Brien and Fleming's type is:

$$f(t_j) = 2\{1 - \Phi(Z_{1-\alpha/2}/\sqrt{t_j})\}, \tag{2.48}$$

Hwang et al. (1990) describe flexible spending function as

$$f(t) = \alpha\frac{(1 - exp(-\gamma t_j))}{(1 - exp(-\gamma))}, \tag{2.49}$$

for $t_j = I_j/I_{max}$, representing the proportion of the total information accumulated.

The boundaries created with γ = 1 are similar to the boundaries from the Pocock method, and the boundaries created with γ = -4 or γ = -5 are similar to the boundaries from the O'Brien-Fleming method. The last two scenarios can be visualised in Figure 2.1.

To achieve a specified power, Jennison and Turnbull (2000a) explain that under $\theta = \delta$, the distribution $\{z_1, ..., z_j\}$ depends on the absolute information levels $\{I_1, ..., I_j\}$. Furthermore, the authors explain that for design purposes, it is necessary to set a maximum number of stage, J, and to assume that the information levels $I_1, ..., I_J$ are specified in advance.

Under the assumption in Eq. (2.45), the value of I_J can be chosen to meet a given power requirement.

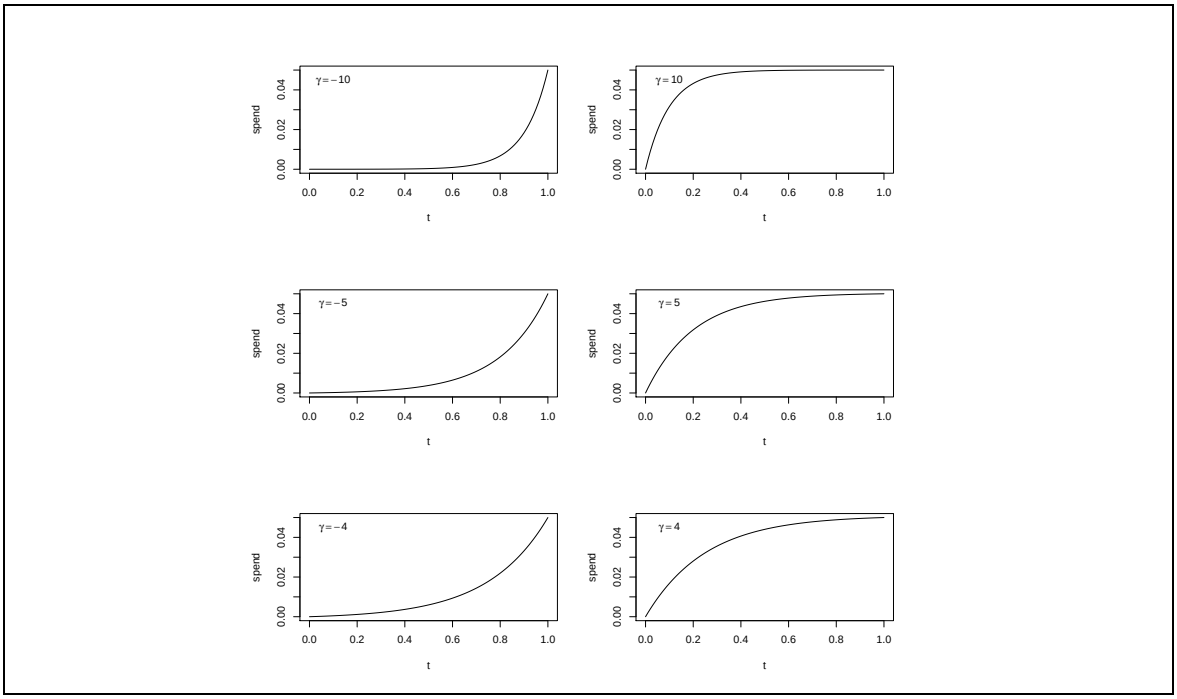

Figure 2.1: Hwang-Shih-DeCani family of type I probability spending functions for various values of γ

2.3.3 Stopping boundary calculation

The critical values $c_1, ...c_J$ satisfying Eq. (2.45) can be found using the method proposed by Armitage et al. (1969) and also described by Jennison and Turnbull (2000a), known as recursive numerical integration. This methodology is used to find the joint distribution of $Z_1, ...Z_J$ when θ = 0 allowing for stopping.

Under H_0, and considering the distribution of the test statistics in Eq. (2.38), at the first interim analysis, we have $Z_1 \sim N(0, 1)$. The density of this distribution is denoted by:

$$f_1(Z_1) = \phi(Z_1) \tag{2.50}$$

where ϕ denotes the standard normal density function. To satisfy Eq. (2.45) for j=1, critical value c_1 are set to the upper π_1 points for the above normal distribution.

For the second interim analysis, the subdensity of Z_2 for those trials that continue to the second stage is given by

$$f_2(Z_2) = \int_{-\infty}^{c_1} f_1(Z_1)\phi(Z_2 - Z_1)dZ_1. \tag{2.51}$$

knowing that increment $Z_2 - Z_1$ is normally distributed and independent of Z_1.

This subdensity can be calculated by evaluating the integral numerically given the value of c_1 found previously. The critical value c_2 to satisfy Eq. (2.39) is then given by the upper π_2 point for this subdensity.

If we continue in this way, the subdensity $f_j(Z_j)$ for Z_j, for those trials that continue to stage *j*, is given recursively

$$f_j(Z_j) = \int_{-\infty}^{c_{j-1}} f_{j-1}(Z_{j-1})\phi(Z_j - Z_{j-1})dZ_{j-1} \tag{2.52}$$

allowing calculation of c_j for all j=1,2...,J to satisfy Eq. (2.39) as required.

2.3.4 Post-trial analysis

At the end of a sequential trial, an analysis must be performed. The interim analyses serve only to determine whether stopping should take place, but they do not provide complete interpretations of the data. The final analysis of the data expresses the degree of evidence that a difference between the experiment and the control group exists using a P-value, and to estimate its magnitude using point estimate and confidence interval (Whitehead (1999)). In this thesis, it will be assumed that the trial has stopped according to a formal stopping procedure.

2.4 Group Sequential Inverse Normal combination tests with a single endpoint

This section describes in more details a method integrating the Inverse Normal combination test approach described in Section 2.2 into the classical Group Sequential Designs illustrated in Section 2.3. This implies using the sample size estimated with the GSD method and the inverse normal combination test statistics to control the type I error rate. Chapter 5 will present the group sequential inverse normal combination tests in the context of multiple co-primary endpoints based on the methodology and notation used in this section. Subsection 2.4.1 presents an introduction, Subsection 2.4.2 describes elements of inverse normal combination test methodology in GSD setting, following by an illustration of stopping rules in subsection 2.4.3.

2.4.1 Introduction

In Subsection 2.2.1, we explained that combination test approaches, as initially developed by Bauer (1989), are based on the principle that the test statistics are calculated separately from different stages data. The test decision is derived from a predefined function that combines the test statistics into a single criterion after each stage. The advantage of the combination test methods is their flexibility that allows the adaptation of certain experimental conditions, such the sample size, the test statistic, or even the outcome variable used to measure the treatment effect, to the data observed in the previous stages of the trial. The advantage of group sequential tests is that there is a range of possible choices for the critical boundaries (Muller and Schafer (2001)). For example, in the planning phase, the group sequential test may optimally be fitted to the clinical situation and the special research problem. One may choose a design out of a number of different plans published in the literature, such as Pocock (1977), O'Brien and Fleming (1979), or Lan and DeMets

(1983), or may construct boundaries that meet the requirements of the individual trial, using fairly simple methods of numerical integration.

However, combination test approaches are based on special combination rules for *p*-values such as Fisher's rule or inverse normal combination test rule and do not offer such a large variety of possible plans as group sequential designs do. On the other hand, group sequential designs do not offer the flexibility to make data adaptive changes to the trial design during the course of the trial based on the results of interim analyses (Muller and Schafer (2001)). That is why, several authors, such as Cui et al. (1999), Lehmacher and Wassmer (1999) and Muller and Schafer (2001), proposed a method that uses the classical stopping boundaries while enabling an adaptive planning of the ongoing trial. The first authors implemented a valid inference procedure that allows flexibility for adjusting sample size based on the updated estimate of treatment effect during the course of the trial. The proposed approach is a group sequential test procedure with pre-specified weights used in the traditional repeated significance two-sample mean test. The second authors proposed a method for group sequential trials based on the inverse normal method for combining the results of the separate stages and the last authors suggested a method for integrating the concept of adaptive interim analysis into classical group sequential testing.

In the following subsection we focus on the method proposed by Lehmacher and Wassmer (1999) which implies that if, at some stage *j*, one always uses the unweighted mean of the test statistics and the critical values designed for the case of equal sample sizes between the stages, then the resulting group sequential test procedure is independent of these sample sizes as long as the test statistics are independent and standard normally distributed. These normal scores are obtained by the inverse normal method, which is a common method in combining test results.

2.4.2 Representing GSD tests as a Combination rule of *j* independent *p*-values

The idea behind the method in this subsection is that we describe a group sequential design defined implicitly as a combination of inverse normal tests of *p*-values, that is, at the time of the first interim analysis, the decision rule of a group sequential test can be represented as a combination rule for *j p*-values, one of the *j p*-values being derived from the data collected in the first stage of the trial and the other *p*-value being derived from the independent data collected in the further course of the trial, i.e., in stages 2 to *j*. At each time point of the interim analysis, the further stage of the trial can be understood as an independent new trial.

To define GSD inverse normal tests, let B defined in Eq. (2.25) now be B_j, expressing the inverse normal test statistic at stage *j*, that is

$$B_j = w_1\Phi^{-1}(1-p_1) + \ldots + w_j\Phi^{-1}(1-p_j)$$

satisfying

$$\sum_{j=1}^{J} w_j^2 = 1.$$

If the weight w_j, $j = 1, \ldots, J$ are determined in advance, the terms $w_j\Phi^{-1}(1-p_j)$, $j = 1, \ldots, J$, are independent, so under the null hypothesis, B_j is normal distributed with mean 0 and variance $w_1^2 + \ldots + w_j^2$, that is

$$B_j \sim N(0, w_1^2 + \ldots + w_j^2).$$

If the p_j are independent and uniform, the covariance between the inverse normal test statistics at stages *j* and $j+1$ is:

$$Cov(B_j, B_{j+1}) = w_j. \tag{2.53}$$

Under H_0, $B_1, B_2, \ldots, B_J$, have a multivariate normal distribution:

$$\begin{pmatrix} B_1 \\ . \\ . \\ . \\ B_J \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ . \\ . \\ . \\ 0 \end{pmatrix} , \begin{pmatrix} w_1 & w_1 & . & . & . & w_1 \\ w_1 & w_2 & & & & w_2 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ w_1 & w_2 & . & . & . & w_J \end{pmatrix} \right). \tag{2.54}$$

Because each *p*-value is computed from a separate dataset, B_j and $(B_{j+1} - B_j)$ are independently distributed and under the joint null distribution, the process B_j is of the form defined Eq. (2.54); consequently when monitored according to the GSD, type I error specifications will be maintained.

2.4.3 Stopping rules

The distribution of the test statistics introduced in Eq. (2.54) are used at each interim stage of a GSD and have this form of the stopping rule:

$$\begin{array}{ll} \text{After stage } j = 1, ..., \text{J-1} & \\ \quad \text{if } B_j \geq c_j & \text{stop, reject } H_0 \\ \quad \text{otherwise} & \text{continue to stage } j+1 \\ \text{after stage J} & \\ \quad \text{if } B_J \geq c_J & \text{stop, reject } H_0 \\ \quad \text{otherwise} & \text{stop, accept } H_0. \end{array} \tag{2.55}$$

where c_j represents a critical value or boundary for Z_j defined in Eq. (2.38). As in the GSD, the boundary c_j is computed exactly as in Subsection 2.3.3.

A type I error probability is:

$$\sum_{j=1}^{J} Pr(B_1 < c_1, \ldots, B_{j-1} < c_{j-1}, B_j \geq c_j | \theta = 0) = \alpha \tag{2.56}$$

for some *j* = 1,...,J, estimated when $\{B_1, \ldots, B_j\}$ follow the null distribution of Eq. (2.54).

The type I error probability can also be expressed in term of spending function as defined in Eq. (2.45), that is:

$$Pr(B_1 < c_1, \ldots, B_{j-1} < c_{j-1}, B_j \geq c_j | \theta = 0) = \pi_j \tag{2.57}$$

where π_j represents the probability of stopping at stage *j* to reject H_0 when this hypothesis is true. It is also called the error spent at stage *j*. It is obtained by specifying c_j calculated using Eq. (2.45).

As explained by Whitehead (2010), the specification and interpretation of alternative hypotheses is more difficult to define in general. However, this thesis considers using the sample size of the GSD method and the GSD inverse normal combination test statistics to maintain the power. This will be checked by simulations in the setting of multiple endpoints.

2.4.4 GSD Inverse Normal combination test: Motivation

This subsection is an extension of Subsection 2.2.4.2 in the setting of J stages. It shows analytically that combination test method in GSD setting maintains the type I error rate for

any possibly data-driven choice of sample sizes. To illustrate this, we begin by extending *Result 1* described in Subsection 2.2.4.2 in *J*-stage setting .

Result 2 (J-stage combination functions): *Result 1* (given by Proschan (2009b)) is true with J stages provided that for each j = 1,. . . ,J, the sample size for stage j depends only on data from stages 1,. . . ,j-1. If this holds, and if $p_1, \ldots, p_j$ are uniformly distributed on (0,1) in a fixed sample setting, then they remain uniformly distributed on (0,1) in the adaptive sample size setting. Therefore, any level α rejection region in a fixed sample setting remains level α in an adaptive setting.

The following example given by Proschan (2009b) explains that *Result 2* can be used to combine monitoring with sample size re-estimation: without sample size modification, we might plan to monitor the data j times over the course of the trial. The stage-specific *B*-scores $B_1, \ldots, B_j$ are iid N(0, 1) under the null hypothesis, and the *B*-score for the cumulative data up to look *j* is

$$B_j = \frac{1}{\sqrt{J}}(B_1 + \ldots + B_j). \tag{2.58}$$

Suppose that $c_1, \ldots, c_j$ is any boundary for $B_1, \ldots, B_j$ in a fixed sample size setting. Then even if we modify sample sizes of subsequent stages on the basis of past data, Proschan (2009b) explains that the stage-specific *B*-scores remain iid N(0,1) under the null hypothesis, and therefore $B_1, \ldots, B_j$ has the same joint distribution in the adaptive sample size setting. Therefore, as long as we apply the boundaries c_j to the cumulative *B*-scores defined in Eq. (2.58) the type I error rate will be α. This is the idea underlying a method proposed independently by Cui et al. (1999) and Lehmacher and Wassmer (1999).

2.5 Summary

This chapter provides a literature review of the methods in the context of a single endpoint. Section 2.1 provides a background on the sample size re-estimation procedure with a single endpoint, Section 2.2 describes the inverse normal combination test method in the context of a single endpoint, Section 2.3 describes the group sequential design method, again in the setting of analysis of a single endpoint and Section 2.4 presents the group sequential design inverse normal combination test method with a single endpoint. The next chapter describes how the SSR approach with a single endpoint, developed in Section 2.1 and Section 2.2, can be extended to the setting of multiple co-primary endpoints.

Chapter 3

Methods for sample size re-estimation with multiple co-primary endpoints without early stopping

This chapter describes how the SSR approach with a single endpoint, developed in Section 2.1 and Section 2.2, can be extended to the setting of multiple co-primary endpoints. Section 3.1 presents SSR method in the context of multiple co-primary endpoints and Section 3.2 illustrates SSR inverse normal combination test approach with multiple co-primary endpoints.

3.1 Sample Size Re-estimation with Multiple Co-primary Endpoints

In the introduction, we explained that the aim of this thesis is to answer two questions. First, how to adjust a sample size in a clinical trial with multiple continuous co-primary endpoints using adaptive and group sequential designs. Second, how to construct a test in such a way to control the FWER and maintain the power, even if the correlation ρ between endpoints is not known. To answer these questions, the following method is proposed: K different tests are conducted, each for one endpoint and each at level α/K, and SSR is performed in which the results of the interim analysis are used to estimate one or more

nuisance parameters. This information is used to determine the sample size for the rest of the trial.

In this section, we start by defining an analysis framework in Subsection 3.1.1, followed by a formulation of the problem for K co-primary endpoints in Subsection 3.1.2. Subsection 3.1.3 presents the construction of *K* tests that satisfy FWER conditions and power requirements, Subsection 3.1.4 illustrates how to calculate sample size in the multiple co-primary endpoints setting, Subsection 3.1.5 describes the implementation of the method, Subsection 3.1.6 presents a worked example and Subsection 3.1.7 presents simulation results.

3.1.1 Framework for the analysis of a SSR with multiple co-primary endpoints

The framework of the analysis developed in Subsection 2.1.2 is extended as follows in the context of multiple co-primary endpoints:

(Step 1) The initial sample size calculation leading to a provisional sample size N_0 is carried out on the basis of an initial estimate of the nuisance parameters $\rho_{kk_0'}$ ($k' > k$) and $\sigma^2_{k_0}$. Further details on how to compute the sample size are given in Subsection 3.1.4.

(Step 2) When the data for the first $n_1 = \pi N_0$ patients (e.g., π = 0.5) are available, then the nuisance parameters $\rho_{kk_1'}$, $k' > k$ and $\sigma^2_{k_1}$ are re-estimated from these observations, which constitute the internal pilot study. These estimates are then used to calculate the sample size N. The sample size can then be adjusted following a predefined recalculation rule:

– restricted design, i.e. $n_2 = max(N_0, N) - n_1$ or

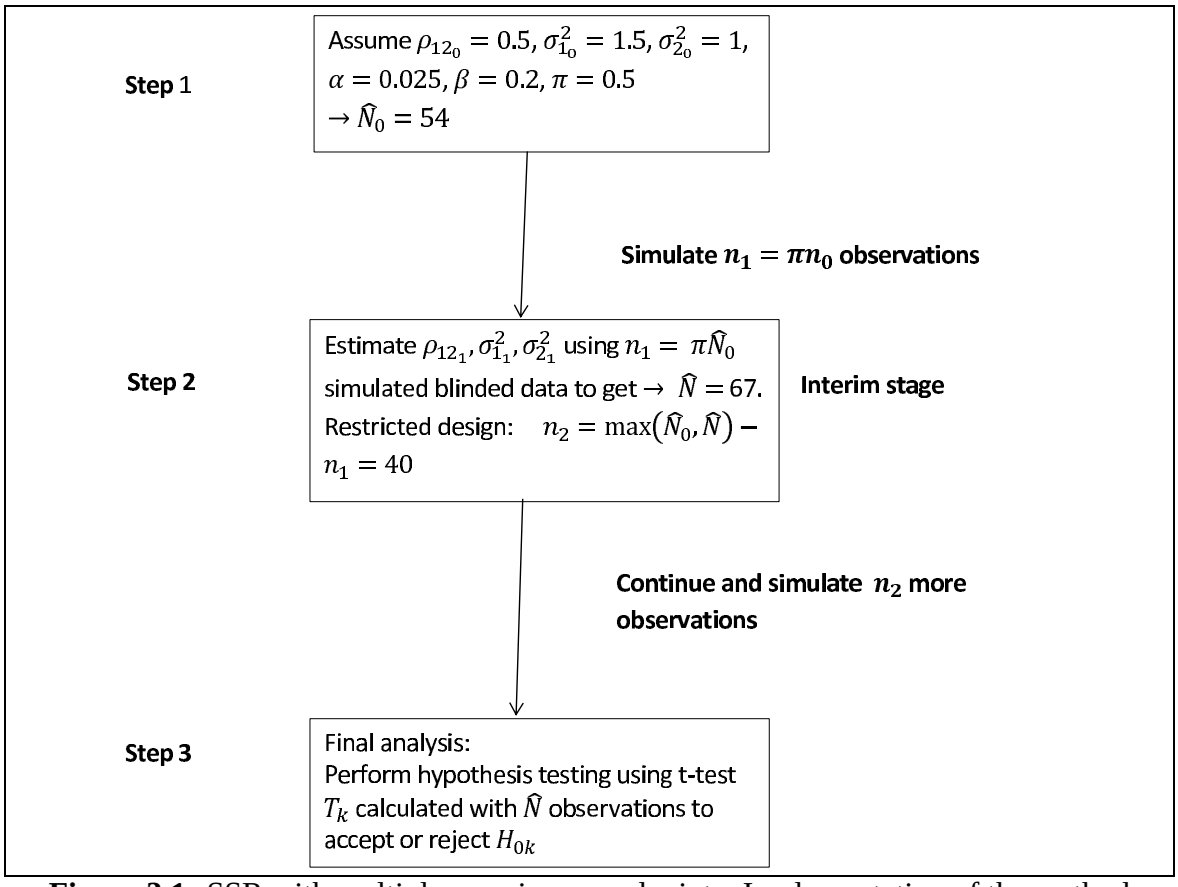

Figure 3.1: SSR with multiple co-primary endpoints: Implementation of the method

– unrestricted design, i.e. n_2 = max(n_1, N) - n_1

This also can be done using:

– unblind method, i.e. use pooled estimate of $\rho_{kk'}$ and σ^2_k

– blind method, i.e. use whole variance or assume difference between groups of δ

(Step 3) The final analysis includes all $N = n_1 + n_2$ observations. The hypothesis test is conducted using a standard t-test.

This framework can also be visualised in Figure 3.1, and the program in Appendix A follows steps of the framework to stimulate the data, calculate the critical value, estimate the sample size and perform the test of hypotheses. More details about the values of each variable in Figure 3.1 are given in Table 3.1. The same values have been used in the worked example described in Subsection 3.1.6.

3.1.2 Formulation of the problem

In this section, we consider methodology for situations where there are K co-primary continuous correlated endpoints in a clinical trial. The general setting for this problem is defined in Subsection 1.4.1. Suppose that E and C are two treatments to be compared in a randomised (phase III) parallel group clinical trial. After each group of N subjects has been randomised in equal numbers to the two therapies and the response obtained, $\rho_{kk'}$ and σ_k^2 are re-estimated and the accumulated data tested. The primary trial's objective is to determine whether E is more efficacious than C in terms of K continuous co-primary responses. This procedure is conducted at the final step of SSR analysis framework described in Subsection 3.1.1, this includes all $N = n_1 + n_2$ observations, which involves a comparison of the evidence of efficacy of E and C, with the rejection occurring as soon as one of the K-hypotheses is in some sense sufficiently convincing.

3.1.3 Test statistics

In this Subsection, we are interested in constructing test statistics in the setting of K co-primary endpoints and deriving their distribution.

Suppose we use the standardised statistic defined in Eq. (1.12) and the information for θ defined in Eq. (1.13) to construct test statistics, and we assume σ_k^2 is known. Let Z_k, k = 1,..,K now denote the standardised statistic for θ_k and endpoint k, which we write as:

$$
\begin{aligned}
Z_k &= \frac{1}{\sqrt{(2N\sigma_k^2)}}(\sum_{i=1}^{N} X_{ikE} - \sum_{i=1}^{N} X_{ikC}) \\
&\sim N((\theta_{kE} - \theta_{kC})\sqrt{\{N/(2\sigma_k^2)\}}, 1)
\end{aligned} \tag{3.1}
$$

and the information is now defined as

$$
I_k = \frac{N}{2\sigma_k^2}. \tag{3.2}
$$

Under H_{0k}, Z_k is normal distributed with mean 0 and variance 1, i.e. $Z_k \sim N(0,1)$.

Suppose $\rho_{kk'}$ is the correlation between endpoints:

$$
Cov(Z_k, Z_{k'}) = \rho_{kk'}, k^{'} > k. \tag{3.3}
$$

Z_k (k = 1,. . ., K) has a multivariate normal distribution:

$$
\begin{pmatrix} Z_1 \\ \vdots \\ Z_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1\sqrt{I_1} \\ \vdots \\ \theta_K\sqrt{I_K} \end{pmatrix} , \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ & 1 & \cdots & \rho_{2K} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right) \tag{3.4}
$$

Replacing I_k by its value in Eq. (3.4), we have

$$
\begin{pmatrix} Z_1 \\ \vdots \\ Z_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1\sqrt{N/2\sigma_1^2} \\ \vdots \\ \theta_K\sqrt{N/2\sigma_K^2} \end{pmatrix} , \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ & 1 & \cdots & \rho_{2K} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right) \tag{3.5}
$$

3.1.3.1 Implications for the FWER

We now need to show that using the critical value c defined in Eq. (1.15), the distribution of the tests constructed in Eq. (3.5) controls the FWER in the strong sense.

So, for one endpoint, in place of Eq. (1.14), we now define

$$\begin{aligned} &Pr(\text{reject } H_0 \mid \theta = 0) = \\ &Pr(Z \geq c) | \theta = 0) = \alpha / K \end{aligned} \tag{3.6}$$

where α/K now describes the error rate adjusted using the Bonferonni correction.

In the setting of K endpoints and using the decision rule defined in Subsection 1.4.1 (vi), in place of Eq. (3.6), we now have:

$$\begin{aligned} &Pr(\text{reject at least one } H_{0k} \mid \theta_k = 0) = \\ &Pr(Z_1 > c \text{ or,...,or } Z_K > c \mid \theta_k\text{=0}) \leq \alpha. \end{aligned} \tag{3.7}$$

So, to control the FWER, one must use c to satisfy Eq. (3.7).

3.1.3.2 Implications for the power

The power is described as in Eq. (1.16), but here we use K multiple co-primary endpoints. So, in place of Eq. (1.16) the power is now given by:

$$\begin{aligned} &Pr(\text{reject at least one } H_{0k} \mid \theta_k = \delta_k) = \\ &Pr(Z_1 > c \text{ or,...,or } Z_K > c \mid \theta_k = \delta_k) = 1 - \beta. \end{aligned} \tag{3.8}$$

For given the values of α, β and c, the sample size can be found that satisfies Eq. (3.8) when Z_k follows the distribution of Eq. (3.5). We use mvtnorm package in R to compute multivariate normal probabilities as described in more details in the next subsection.

3.1.4 Sample size calculation

The aim of this subsection is to show how to calculate the sample size of a design with multiple co-primary endpoints. As stated in the previous subsection, given the values of α/K, β and c, a sample size can be found that satisfies Eq. (3.8) when $Z_k, k = 1, \ldots, K$ follows the distribution of Eq. (3.5). This is done by using the ***mvtnorm*** package in R to compute the multivariate normal probability of the following equation:

$$\Phi(c, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^K}} \int_{-\infty}^{c} \int_{-\infty}^{c} ... \int_{-\infty}^{c} e^{-\frac{1}{2}x^t \Sigma^{-1} x} dX \tag{3.9}$$

where $X = (x_1, x_2, ..., x_K)^t$, $-\infty < c < +\infty$ and Σ is a KxK semi-definite symmetric covariance matrix.

Within the mvtnorm package, the ***pmvnorm***(lower, upper, mean, corr) function is used to compute such probability. The lower argument of the ***pmvnorm*** function represents the vector of lower limits of length K. In the one sided setting, it is

$$\text{Lower limits} = \begin{pmatrix} -\infty_1 \\ \vdots \\ -\infty_K \end{pmatrix}. \tag{3.10}$$

The upper argument indicates the vector of upper limits of length K, that is

$$\text{Upper limits} = \begin{pmatrix} c \\ \vdots \\ c \end{pmatrix}. \tag{3.11}$$

The mean argument represents the mean vector of length K, that is

$$\text{Mean vector} = \begin{pmatrix} \theta_1 \sqrt{\frac{N}{2\sigma_1^2}} \\ \vdots \\ \theta_K \sqrt{\frac{N}{2\sigma_K^2}} \end{pmatrix}. \tag{3.12}$$

The correlation argument indicates the variance-covariance matrix of dimension K, that is:

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ & 1 & \cdots & \rho_{2K} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}. \tag{3.13}$$

The sample size *N* is then found iteratively using the function ***uniroot*** in R, which searches the interval from lower to upper for a root (i.e., zero) of the ***pmvnorm***(lower, upper, mean, corr) function with respect to its first arguments. The lower and upper arguments of the uniroot function are the end points of the interval to be searched.

3.1.5 Implementation of the method

This section illustrates how the framework of analysis described in subsection 3.1.1 could be implemented in practice. It is based on the extension of a three step procedure developed by Wittes and Brittain (1990). The following is an illustration of such a method:

3.1.5.1 Step 1 - Initial sample size calculation

In step 1, the initial sample size calculation leading to a provisional sample size N_0 is carried out on the basis of an initial estimate of the nuisance parameters.

A1.1. Guess $\rho_{kk_0'}$ ($k' > k$) and $\sigma^2_{k_0}$; or consider initial estimates of studies in earlier phases of the drug development process.

A1.2. Determine α and target power = 1 - β.

A1.3. Calculate the critical value c as in Eq. (1.15).

A1.4. Calculate the initial sample size N_0 as illustrated in more detail in Subsection 3.1.4.

A1.5. Fix the fraction of the initial sample size π to be used at interim step.

3.1.5.2 Step 2 - Sample size re-estimation

In step 2, the nuisance parameters are re-estimated based the interim data collected, which constitute the internal pilot study. These estimates are then used to calculate the sample size N for the remainder of the trial.

A2.1. Simulate $n_1 = \pi N_0$ observations.

A2.2. Estimate $\sigma^2_{k_1}$ using blinded method (as in Eq. (2.10)) based on n_1 observations.

A2.3. Use n_1 observation to estimate $\rho_{kk_1'}$ as in Eq. (1.9).

A2.4. Use α and power defined in step (A1.2.), *c* defined in step (A1.3.), $\sigma^2_{k_1}$ and $\rho_{kk_1'}$ estimated in steps (A2.2.) and (A2.3.) respectively to re-calculate the initial sample size N as illustrated in Subsection 3.1.4.

A2.5. Collect n_2 using the restricted design which requires a final sample size at least as large as the original calculated as in Eq. (2.7) .i.e. $n_2 = N - n_1$, where $N = max(N_0, N)$.

3.1.5.3 Step 3 - Final analysis

In step 3 which includes all $N = n_1 + n_2$ observations, the hypothesis test is conducted using a standard *t*-test.

A3.1. Estimate $\sigma^2_{k_2}$ using blinded method (as in Eq. (2.10)), based on all $N = n_1 + n_2$ observations.

A3.2. Use $\sigma^2_{k_2}$ estimate in step (A3.1.) and $N = n_1 + n_2$ to calculate *T*-statistic as in Eq. (2.6).

A3.3. Calculate the critical value c as in Eq. (2.14).

A3.4. Reject at least one H_{0k} at level α if $T_k \geq c$. Note that T_k is defined in Eq. (1.23) in a single endpoint context but in this setting, it is a *t*-statistic for endpoint k calculated with N observations.

3.1.6 Example: SSR with Multiple Co-primary Endpoints

Suppose that E and C are two treatments to be compared in a randomised (phase III) parallel group clinical trial. Two co-primary endpoints are considered, i.e. K = 2. Patients are randomised in equal numbers between E and C, and a normally distributed response is observed for each of the endpoints. Suppose the parameters of interest representing the mean differences are $\theta_1 = \theta_2 = 0.5$.

Table 3.1: SSR: Implementation of the method

Steps	**Values**
Step 1	
Significance level α	0.025 (one sided)
Power $1-\beta$	0.8
Endpoints	K = 2
Assume ρ_{12_0}	0.5
Assume $\sigma^2_{1_0}$	1.5
Assume $\sigma^2_{2_0}$	1
Assume π	0.5
Calculate N_0	54
Step 2	**Interim step**
Simulate $n_1 = \pi N_0$ data	27
Estimate ρ_{12_1}	0.53
Estimate $\sigma^2_{1_1}$	1.30
Estimate $\sigma^2_{2_1}$	0.94
Estimate N	67
Use restric. design: n_2	40
Simulate n_2 data	40
Step 3	**Final analysis**
Calculate critical value c with N data	1.98
Calculate T_1 and T_2 with N data	0.84 and 2.42
Conclusion $T_1 < c$ and $T_2 > c$	Stop, reject H_{0k}

In step 1 (see Subsection 3.1.5.1 and Figure 3.1), the values considered are summarized in Table 3.1. We assume (or guess) that the variance for endpoint 1 is $\sigma^2_{1_0} = 1.5$, the variance for endpoint 2 is $\sigma^2_{2_0} = 1$ and the correlation between endpoints is $\rho_{12_0} = 0.5$. A SSR with multiple co-primary endpoints is required to test $H_{0k} : \theta_k$ = 0, k = 1,2, with a one-sided test type I error rate of $\alpha = 0.025$ and a power of 1 - β = 0.80 for $\theta = 0.5$. We use step (A1.3.) to calculate the critical value, that is $c = \Phi^{-1}(1 - \alpha/K)$ which can be evaluated in R using ***qnorm***(1-0.025/2) = 2.241403. We then use step (A1.4.) to estimate the initial sample size N_0 by defining the mean vector, variance-covariance matrix, the lower and upper limit vectors respectively as follow

$$\begin{pmatrix} \theta_1\sqrt{\frac{N}{2\sigma^2_{1_0}}} \\ \theta_2\sqrt{\frac{N}{2\sigma^2_{2_0}}} \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12_0} \\ \rho_{12_0} & 1 \end{pmatrix}, \begin{pmatrix} -\infty_1 \\ -\infty_2 \end{pmatrix} \text{ and } \begin{pmatrix} c \\ c \end{pmatrix}.$$

This gives an initial sample size of $N_0 = 54$. We finally fix the fraction of the initial sample size to be used at interim step by $\pi = 0.5$.

In step 2, the values simulated and estimated are summarized in Table 3.1. We simulate $n_1 = \pi N_0 = 0.5 * 54 = 27$ observations; then the nuisance parameters $\sigma^2_{k_1}$ and ρ_{12_1} are re-estimated from these observations as in steps (A2.2) and (A2.3) respectively. The initial sample size N_0 is re-calculate by replacing the values of $\sigma^2_{k_1}$ and ρ_{12_1} in the the mean vector and the variance-covariance matrix as in step 2. Suppose this gives $N = 67$. We carry on by simulating n_2 data as in step (A3.1) i.e. $n_2 = N - n_1 = 67 - 27 = 40$.

In step 3, the values calculated and the conclusions of hypothesis testing are summarized in Table 3.1. We estimate $\sigma^2_{k_2}$ as in step (A3.2.). We then calculate *T*-statistic and critical value $c = 1.98$ as in step (A3.3.) and (A3.4.) respectively. We finally perform hypothesis testing by rejecting H_{01} or H_{02} at level α if $T_1 \geq c$ or $T_2 \geq c$.

Table 3.2: Initial values considered in the simulation study.

Fixed parameters	SSR
Significance level α	0.025 (one sided)
Standard error of estimate FWER	0.001
Target power $1-\beta$	0.8
Standard error of estimate power	0.0025
Number of endpoints	K = 2
Number of simulations	100,000
Null hypothesis H_{0k}	$\theta_1 = \theta_2$ = 0
Guessed nuisance parameters	
ρ_{12_0}	0.5
$\sigma^2_{1_0}$	1.5
$\sigma^2_{2_0}$	1

3.1.7 Simulation results

As we said in the introduction of this section, the procedure for sample size re-estimation with continuous data aims to maintain the desired power of the study without inflating the FWER above the nominal level, even if the nuisance parameters $\rho_{kk'}$ and σ^2_k are not known at the planning stage. In this section, we evaluate these characteristics by simulations. We focused on situations that are typical for phase III trials with two co-primary endpoints. Table 3.2 presents the fixed values considered in the simulation study. In all the scenarios to be described in Table 3.3 below, we conduct 100,000 simulated trials (standard error of estimate FWER $\alpha = 0.025$ is 0.001 and 0.0025 for the power 1-β = 0.80). We consider the initial guess correlation ρ_{12_0} to be 0.5, the initial guess variance for endpoint 1 $\sigma^2_{1_0}$ to be 1.5 and for endpoint 2 $\sigma^2_{2_0}$ to be 1. We aim to randomise patients in equal numbers between E and C.

The scenarios considered in Table 3.3 have the following variable values:

In scenario 1, we consider the parameters of interest representing the mean difference to be ($\theta_1 = \theta_2 = 0.5$). We also consider the proportion of interim data π to be 0.50 and

Table 3.3: Scenarios considered in the simulation study.

Variable values	SSR
Scenario 1	
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Proportion π in SSR	0.50
True nuisance parameters	
ρ_{12}	0,0.1,...,1
σ_1^2	1.5
σ_2^2	1
Scenario 2	**Common values for all settings**
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Proportion π in SSR	0.50
True nuisance parameters	
ρ_{12}	0,0.1,...,1
σ_1^2	1,1.1,1.2,...,2
Setting 1	
σ_2^2	1
Setting 2	
σ_2^2	1.2
Setting 3	
σ_2^2	1.5
Setting 4	
σ_2^2	1.8
Setting 5	
σ_2^2	2
Scenario 3	**Constant** ρ_{12}
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Proportion π in SSR	0.50
True nuisance parameters	
ρ_{12}	0.5
σ_1^2	1,1.1,1.2,...,2
σ_2^2	1,1.1,1.2,...,2
Scenario 4	**Different size effects**
Alternative hypothesis	$\delta_1 = 0.5$, $\delta_2 = 0.7$
Alternative hypothesis	$\delta_1 = 0.7$, $\delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	
Scenario 5	**Different proportion** π **in SSR**
Proportion π in SSR	0.10, 0.80
Same alternative hypothesis as in Setting 3	$\delta_1 = \delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	

simulate data with the following true characteristics: ρ_{12} ranged from 0 to 1 .i.e. (0, 0.1,. . . , 1), true pooled variance for endpoint 1 $\sigma_1^2 = 1.5$ and true pooled variance for endpoint $\sigma_2^2 = 1$.

In scenario 2, we consider five settings with the same parameters of interest representing the mean difference to be ($\theta_1 = \theta_2 = 0.5$) and the same proportion of interim data to be 0.50. The five settings have the same values of ρ_{12} ranging from 0 to 1 .i.e. (0, 0.1,. . . , 1) and the same values of true pooled variances for endpoint 1 σ_1^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2). However, each setting has the following true pooled variance for endpoint 2: In setting 1, σ_2^2 = 1; setting 2, σ_2^2 = 1.2; setting 3, σ_2^2 = 1.5; setting 4, σ_2^2 = 1.8; and setting 5, σ_2^2 = 2.

In scenario 3, we consider ρ_{12} to be 0.5, pooled variance for endpoint 1 σ_1^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2) and pooled variance for endpoint 2 σ_2^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2).

In scenario 4, we consider the same true nuisance parameters as in Setting 3 with different parameters of interest representing the mean difference to be ($\theta_1 = 0.5$, $\theta_2 = 0.7$) and ($\theta_1 = 0.5$, $\theta_2 = 0.7$).

In scenario 5, we consider different proportion of interim data to 0.10 and 0.8 respectively, with the same true nuisance parameters and the same parameters of interest representing the mean difference as in Setting 3.

3.1.7.1 Power in the fixed sample size design for the guess values of the nuisance parameters

Figure 3.2 presents the power in the fixed sample size design for two correlated endpoints. We consider the fixed values defined in Table 3.2 and the variable values defined in Table 3.3, scenario 1. We want to estimate the sample size for the guess values of the nuisance

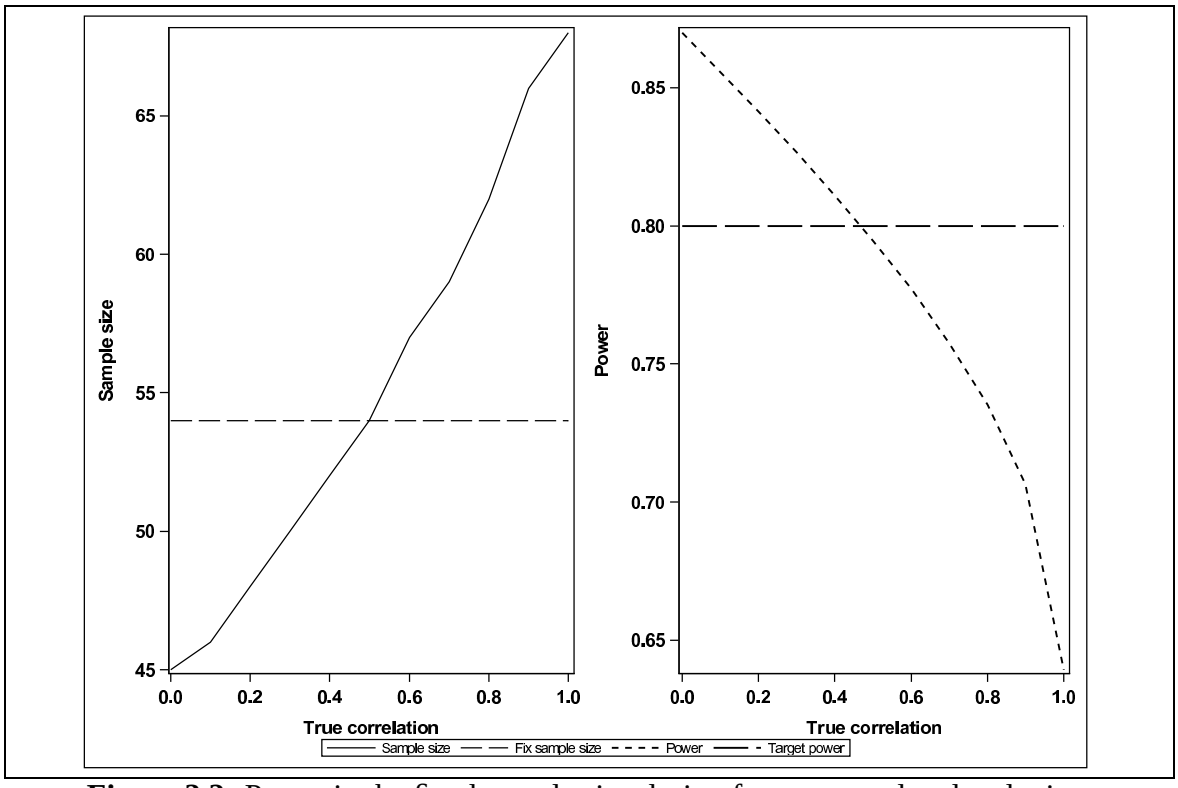

Figure 3.2: Power in the fixed sample size design for two correlated endpoints

parameters without knowing their true values.

The solid line of the left panel in Figure 3.2 indicates the sample size for various values of true correlation $(\rho_{12} = 0, 0.1, \ldots, 1)$, while the dashed line corresponds to the sample size for the guess $\rho_{12_0} = 0.5$. The figure in left panel shows that the sample sizes of the true $\rho_{12} = 0, 0.1, \ldots, 0.4$ are below the one of the guess $\rho_{12_0} = 0.5$. It also shows that the sample size of the true $\rho_{12} = 0.5$ is equal to the one of the guess $\rho_{12_0} = 0.5$, while the sample size of the true $\rho_{12} = 0.6, 0.7, \ldots, 1$ are above the one of the guess $\rho_{12_0} = 0.5$.

In the right panel, the light dashed line shows the power at various points of the true ρ_{12}, while the bold dashed line indicates the target power for the guess $\rho_{12_0} = 0.5$. The figure in the right panel shows that if the guess ρ_{12_0} is above the true ρ_{12} in the left panel, its sample size is above the one of the true ρ_{12}, consequently the power in the left panel is

above the target power, and vice versa; and if the guess ρ_{12_0} is equal to the true ρ_{12}, they generate the same sample size, consequently give the same power.

The results in Figure 3.2 have a number of consequences. At the planning stage of a clinical trial, it is often quite uncertain to know the size of parameters needed for sample size calculation. Figure 3.2 illustrates this. For example, if a clinical trial investigator decide to run a trial with a sample size of 54 above the true sample size of 48 as shown on the left panel of Figure 3.2, this would lead to a more powerful trial than originally planned, with the power of 0.8613 above the target power of 0.80. However, the trial would be unethical as more patients would be exposed to an inferior treatment, and the trial would also cost more money and require more time to be completed. A solution to this problem is to perform a SSR in which the results of the interim analysis are used to re-estimate ρ_{12} and this information is used to determine the sample size for the rest of the trial to make sure it (the trial) is not unnecessarily large for ethical reasons, budget restrictions and time pressure. We present how this could be done in the following subsections.

3.1.7.2 FWER, power and sample size in SSR design

The aim of this subsection is to check the effect of the mis-specification of the nuisance parameters on the FWER, power and sample size using SSR design; in other word, we need to check if the FWER would be controlled and the power maintained if the nuisance parameters change as in the following settings:

3.1.7.2.1 Scenario 2 : FWER in Settings 1 - 5

Figure 3.3 presents SSR FWER's simulation results with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. It shows that the method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12}, σ_1^2 and σ_2^2

in all five settings. In setting 1, the FWER has a minimum value of 0.01101 for perfectly correlated data i.e. $\rho_{12} = 1$ and a maximum value of 0.02264 for uncorrelated data, i.e. $\rho_{12} = 0$. In setting 2, a minimum value of 0.01012 for $\rho_{12} = 1$ and a maximum value of 0.02315 for $\rho_{12} = 0$ have been observed. In setting 3, the FWER has a minimum value of 0.01176 for $\rho_{12} = 1$ and a maximum value of 0.02345 for $\rho_{12} = 0$. In setting 4, a minimum value of 0.01012 for $\rho_{12} = 1$ and a maximum value of 0.02420 for $\rho_{12} = 0$ have been observed. Finally, in setting 5, the FWER has a minimum value of 0.01105 for $\rho_{12} = 1$ and a maximum value of 0.02451 for $\rho_{12} = 0$.

3.1.7.2.2 Scenario 2 : Sample size in Settings 1 - 5

Figure 3.4 presents SSR sample size simulation results with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. The results are presented in all five settings of scenario 2. The figure shows that the sample size increases as ρ_{12}, σ_1^2 and σ_2^2 increase. This is known as the sample size is proportional to the variance. The results in Figure 3.4 also show that the method is working as expected.

3.1.7.2.3 Scenario 2 : Power in Settings 1 - 5

Figure 3.5 presents simulation results for the SSR power with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. The results are presented in all five settings of scenario 2. For example in setting 1, the power has a minimum value of 0.7808 for uncorrelated data i.e. $\rho_{12} = 0$ and a maximum value of 0.7895 for perfectly correlated data i.e. $\rho_{12} = 1$. In setting 2, a minimum value of 0.7801 for $\rho_{12} = 0$ and a maximum value of 0.7994 for $\rho_{12} = 1$ have been observed. In setting 3, the power has a minimum value of 0.7800 for $\rho_{12} = 0$ and a maximum value of 0.7901 for $\rho_{12} = 1$. In setting 4, a minimum value of 0.7810 for $\rho_{12} = 0$ and a maximum value of 0.7900 for $\rho_{12} = 1$

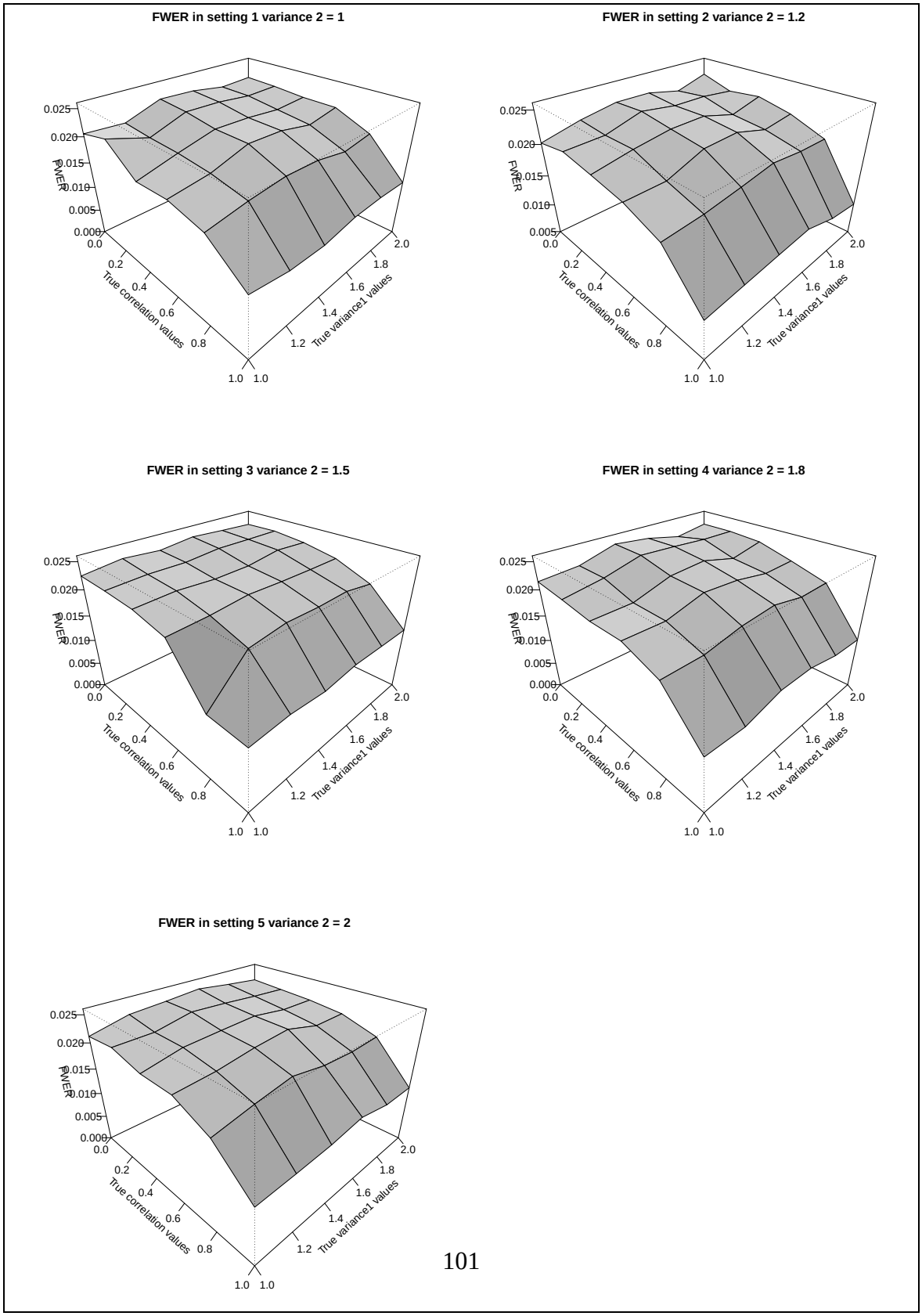

Figure 3.3: SSR FWER in Scenario 2; Settings 1 - 5

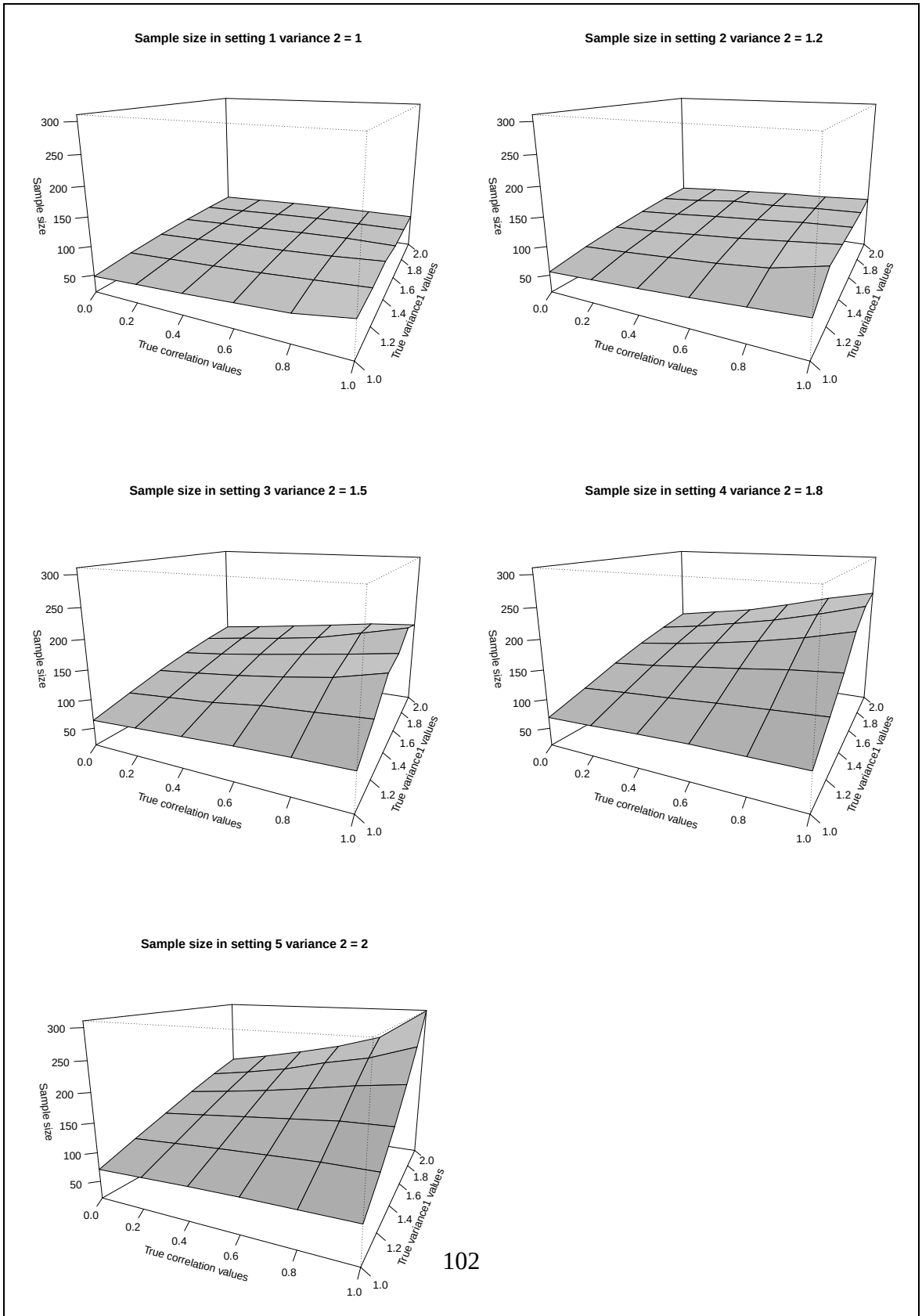

Figure 3.4: SSR Sample size in Scenario 2; Settings 1 - 5

have been observed. Finally, in setting 5, the power has a minimum value of 0.7801 for $\rho_{12} = 0$ and a maximum value of 0.7991 for $\rho_{12} = 1$. Figure 3.5 illustrates that SSR method effectively maintains the power and this is constant despite variation of ρ_{12}, σ_1^2 and σ_2^2. By observing the sample size increasing in the same direction as the variances in Figure 3.4, we would expect the constant power at the nominal level of 0.80. This is an indication that the method is working as expected.

3.1.7.2.4 Scenario 3: Constant ρ_{12}

The results in scenario 3 are presented in Figure 3.6. It (figure) illustrates that despite variation of σ_1^2 and σ_2^2, the FWER is controlled and fairly constant with the minimum value of 0.02079 and the maximum value of 0.02284. The same figure illustrates that the sample size increases in the same direction as σ_1^2 and σ_2^2 with the minimum value 58 and the maximum value 220. Finally the same figure illustrates that the power is maintained and fairly constant despite variation of σ_1^2 and σ_2^2 with 0.7820 the minimum value and 0.7940 the maximum value.

3.1.7.2.5 Scenarios 2 and 3: Summary and comments on the results

The results in scenario 2 show that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases (Settings 1 - 5). The results in scenario 2 also show that the FWER increases as σ_2^2 increases, however, this is true only when the data are not correlated .i.e, $\rho_{12} = 0$ and is assigned to simulation error. The results obtained in this scenario are in line with what we would expect to have because we have used the Bonferroni correction and more details about its characteristics are described in Subsection 1.3.3.1.1. The results in scenario 3 show that the FWER is controlled and fairly constant when ρ_{12} is constant (Scenario 3). Again this is known and this shows that the method is working as expected.

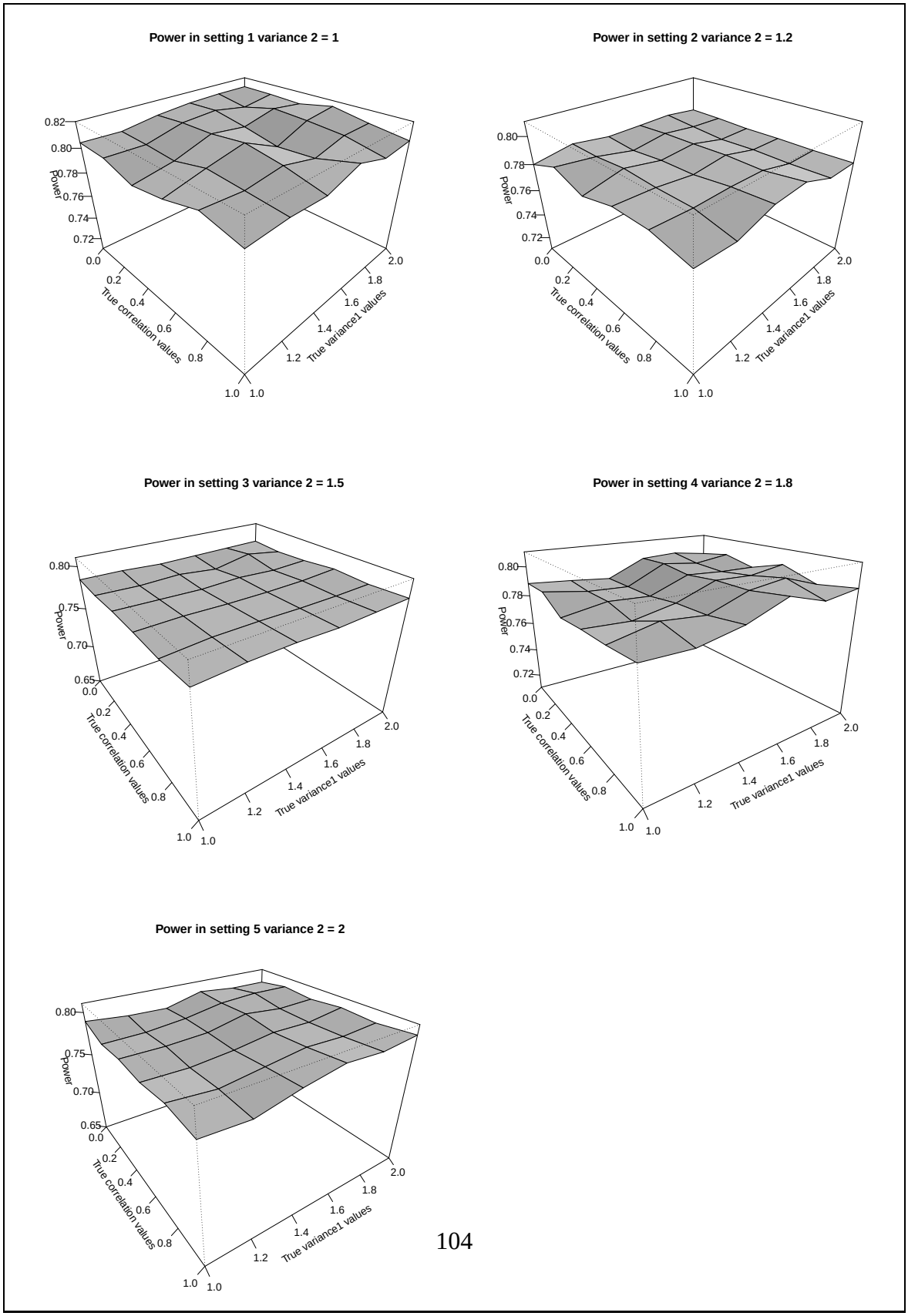

Figure 3.5: SSR Power in Scenario 2; Settings 1 - 5

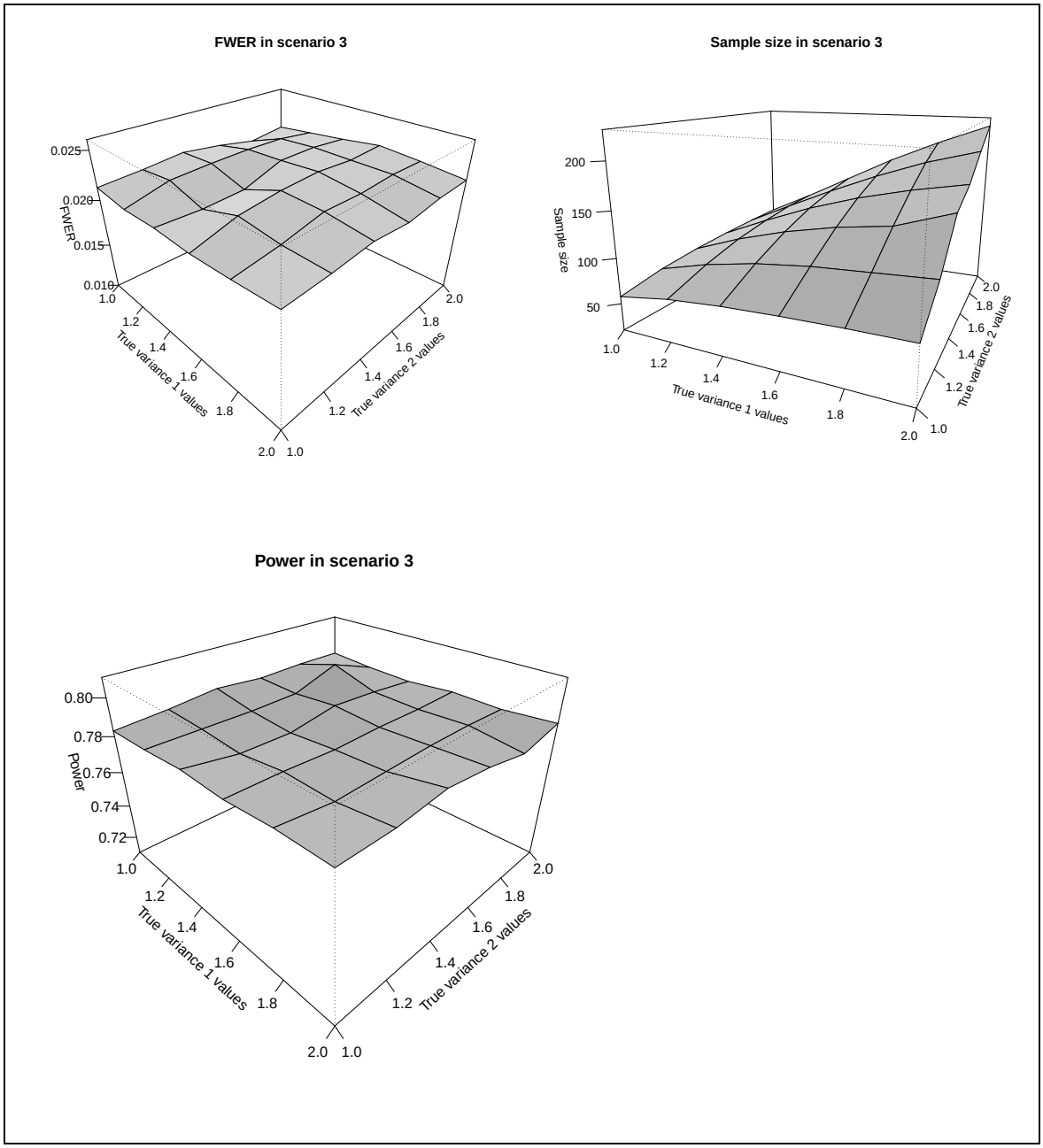

Figure 3.6: SSR FWER, Sample size and Power in Scenario 3

The results in scenario 2 also show that the sample size is increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2, and the power is fairly constant despite variations of these nuisance parameters. This is an indication that this method is adjusting for the sample size needed to maintain the power as defined at the design stage. This is also true for the results in scenario 3.

The results in scenario 2 finally show that all settings control the FWER and maintain the power therefore we only consider Setting 3 to check the characteristics of the FWER, power and sample size when different effect sizes and different timings of the interim analysis are considered.

3.1.7.3 Different effect sizes

3.1.7.3.1 Scenario 4: $\delta_1 = 0.5, \delta_2 = 0.7$

In this subsection, simulations are conducted to check the effect of the variations of the effect size on the FWER, sample size and power. The fixed and variable values considered are described in Table 3.2 and Table 3.3 respectively. Figure 3.7 illustrates a situation where $\delta_1 = 0.5$ and $\delta_2 = 0.7$. It shows that the FWER in this setting is controlled with a minimum value of 0.01123 for $\rho_{12} = 1$ and a maximum value of 0.02324 for $\rho_{12} = 0$. The same figure shows that the sample size is increasing in the same direction as ρ_{12} and σ_1^2 with a minimum value of 51 and a maximum value of 98. Finally, Figure 3.7 shows that the power in this setting is not maintained for the combination of ρ_{12} = (0-0.6) and σ_1^2 = (1-1.2), but maintained for any other combination of ρ_{12} and σ_1^2.

3.1.7.3.2 Scenario 4: $\delta_1 = 0.7, \delta_2 = 0.5$

Figure 3.8 presents a situation where $\delta_1 = 0.7$ and $\delta_2 = 0.5$. It shows that the FWER

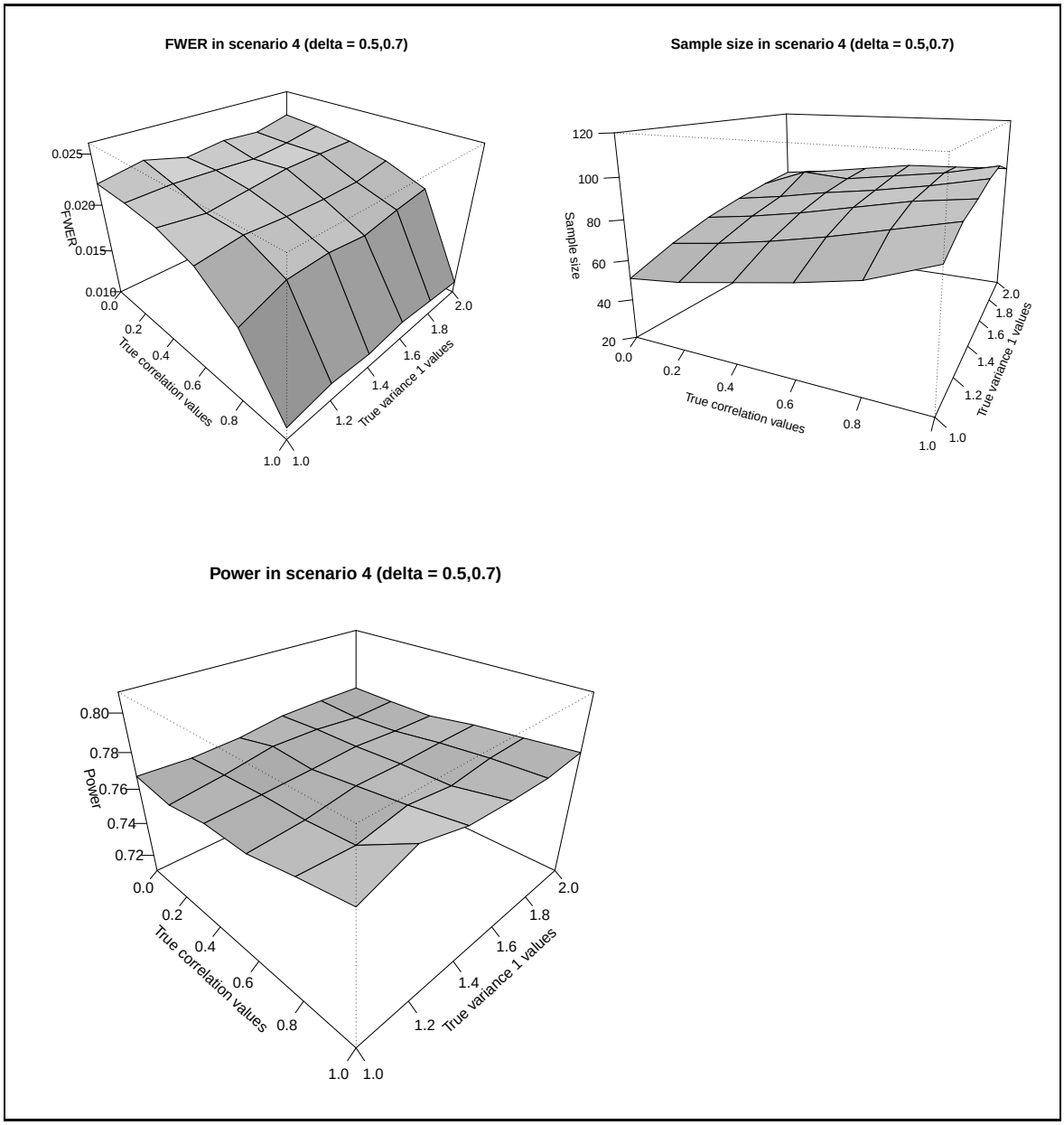

Figure 3.7: SSR FWER, Sample size and Power in Scenario 4 ($\delta_1 = 0.5$, $\delta_2 = 0.7$)

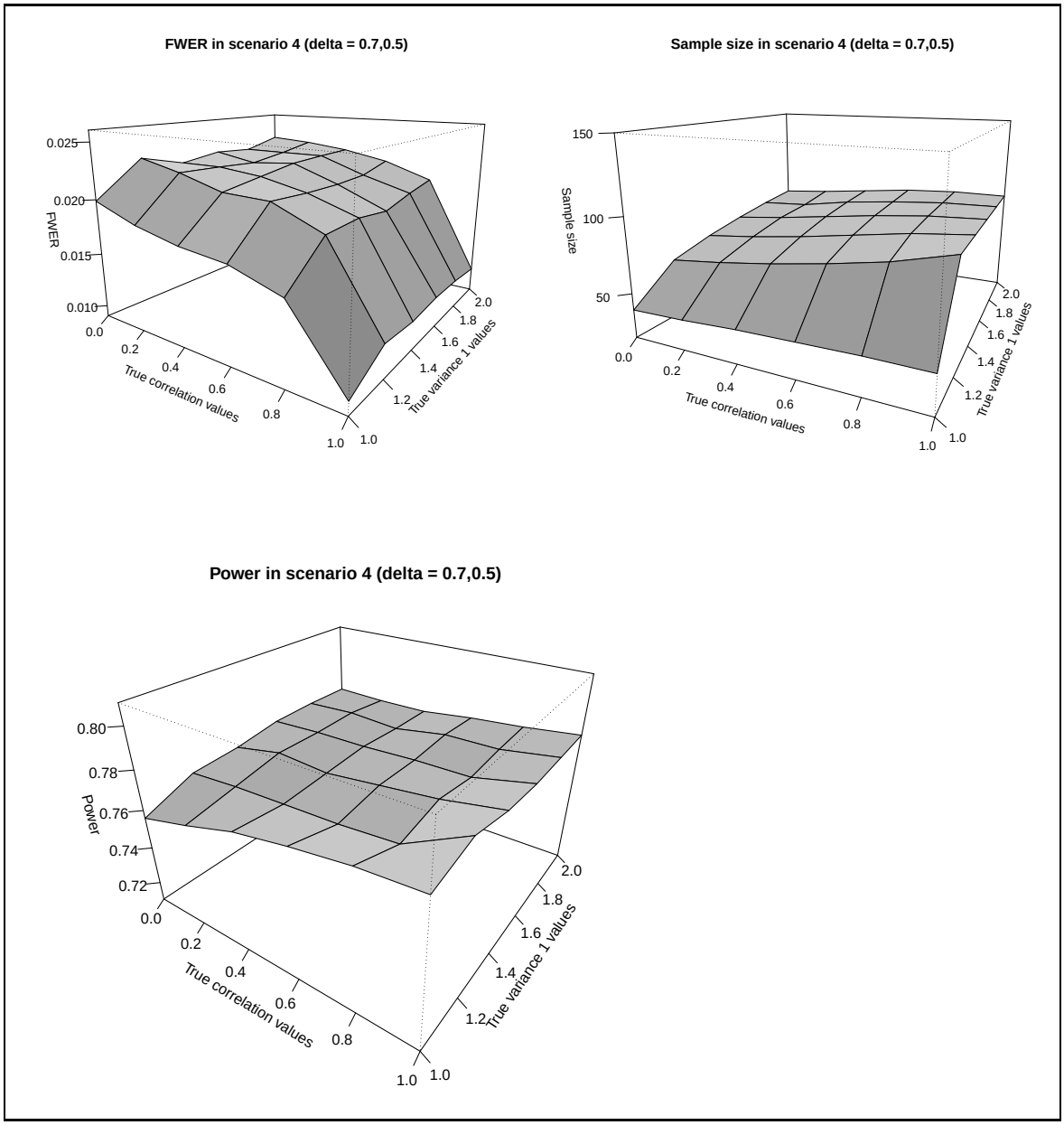

Figure 3.8: SSR FWER, Sample size and Power in Scenario 4 ($\delta_1 = 0.7$, $\delta_2 = 0.5$)

in this setting is controlled with a minimum value of 0.01011 for $\rho_{12} = 1$ and a maximum value of 0.0222 for $\rho_{12} = 0$. The same figure shows that the sample size is increasing in the same direction as ρ_{12} and σ_1^2 with a minimum value of 39 and a maximum value of 92. However, the power in Figure 3.7 is not maintained for the combination of ρ_{12} = (0-0.6) and σ_1^2 = (1-2), but maintained for any other combination of ρ_{12} and σ_1^2.

3.1.7.3.3 Scenario 4: Summary and comments on the results

The results in Scenario 3 show that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases. The results in setting ($\delta_1 = 0.7$ and $\delta_2 = 0.5$) show that the FWER is even more conservative compared to the findings in setting ($\delta_1 = 0.5$ and $\delta_2 = 0.7$), however, this is assigned to simulation error.

The results in Scenario 4 also show that sample sizes are increasing in the same direction as ρ_{12}, σ_1^2, however they (sample sizes) are not large enough to detect different effect sizes at the same time, hence reduction in power. The comment about this scenario is that if different effect sizes are considered in a trial (e.g. $\delta_1 = 0.7$ and $\delta_2 = 0.5$), it is recommended to use the small effect size for sample size calculation. This is a guarantee that the sample size obtained is large enough to detect even the large effect size and maintain the power.

3.1.7.4 SSR: Different timings

3.1.7.4.1 Scenario 5: $\pi = 0.10$

Figure 3.9 presents simulation results with the fixed variable values defined in Table 3.2 and the variable values of Scenario 5 defined in Table 3.3. The situation of $\pi = 0.10$ is illustrated. The figure shows that SSR design controls the FWER with the minimum value

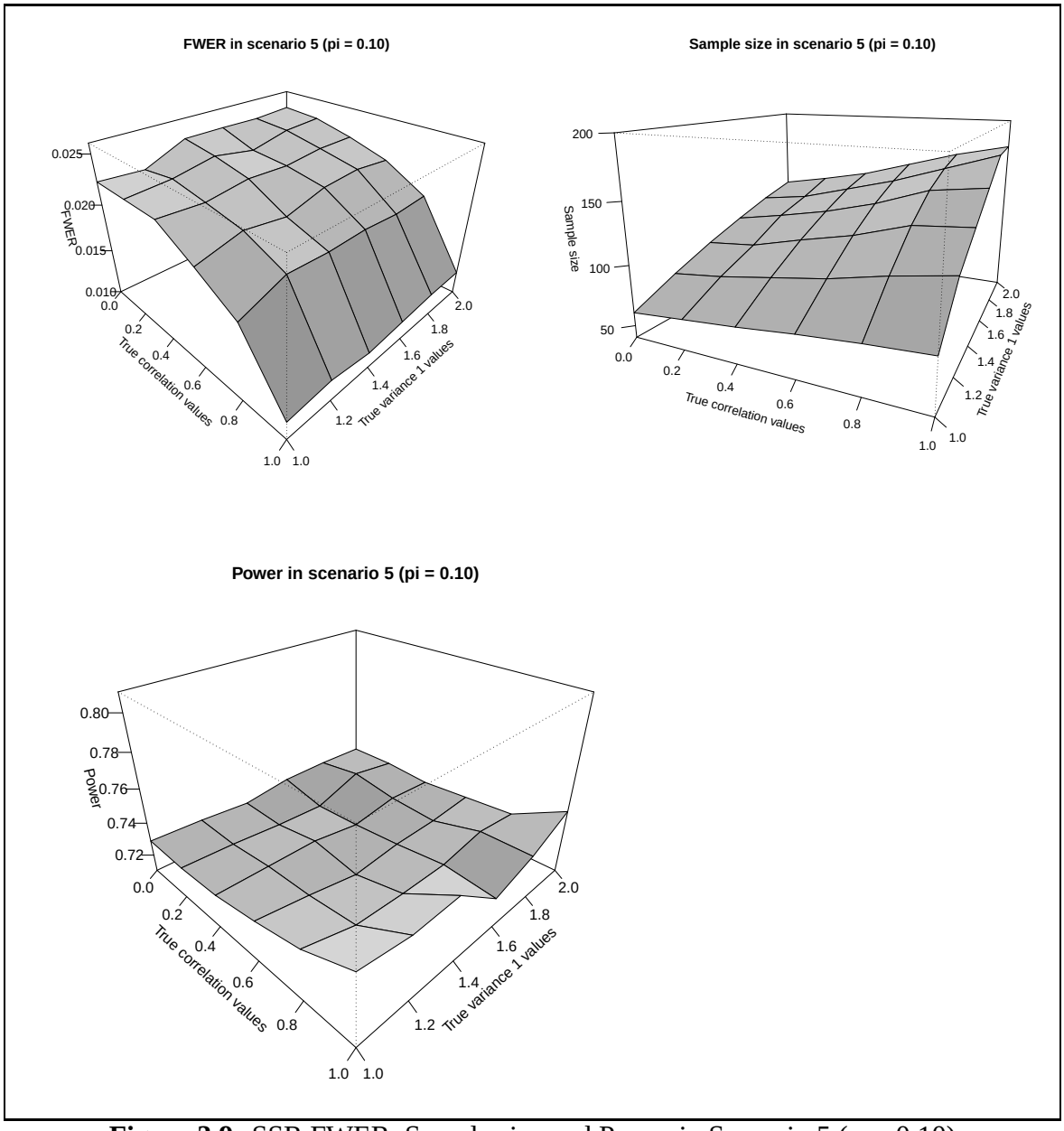

Figure 3.9: SSR FWER, Sample size and Power in Scenario 5 (π = 0.10)

0.01164 for perfectly correlated data i.e. $\rho_{12} = 1$ and 0.02413 for uncorrelated data, i.e. $\rho_{12} = 0$. The same figure also shows that the sample size increases as ρ_{12} and σ_2^2 increase with a minimum value of 61 and maximum value of 176. However, Figure 3.9 shows that, in this setting, the method does not maintain the power when ρ_{12} and σ_2^2 vary. This is due to bias in estimation of ρ_{12} and σ_2^2 from small samples. The minimum and maximum values observed are 0.7240 and 0.7500 respectively.

3.1.7.4.2 Scenario 5: $\pi = 0.8$

Similar to Figure 3.9, Figure 3.10 presents the situation of $\pi = 0.80$. It shows that SSR method effectively maintains the overall FWER at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 with the minimum value 0.01140 for perfectly correlated data i.e. $\rho_{12} = 1$ and 0.02376 for uncorrelated data, i.e. $\rho_{12} = 0$. The figure shows that the method become increasingly conservative as ρ_{12} increases. The same figure also shows that the sample size increases as ρ_{12} and σ_1^2 increase with a minimum value of 64 and maximum value of 176. Finally the same figure shows that the method maintains the power and this is constant despite variation of ρ_{12} and σ_1^2 with a minimum value of 0.7810 and maximum value of 0.7980.

3.1.7.4.3 Scenario 5: Summary and comments on the results

As noted by Friede and Schmidli (2010), the timing of the interim evaluation in clinical trials is not only important for logistic motivations but also affects the operational characteristics of the recalculation procedure. An early interim review cannot give a good estimate of the nuisance parameters, while a very late sample size review may lead to a larger sample size than needed. These scenarios are illustrated in scenario 5.

The results in scenario 5 show that the FWER is controlled but becomes increasingly

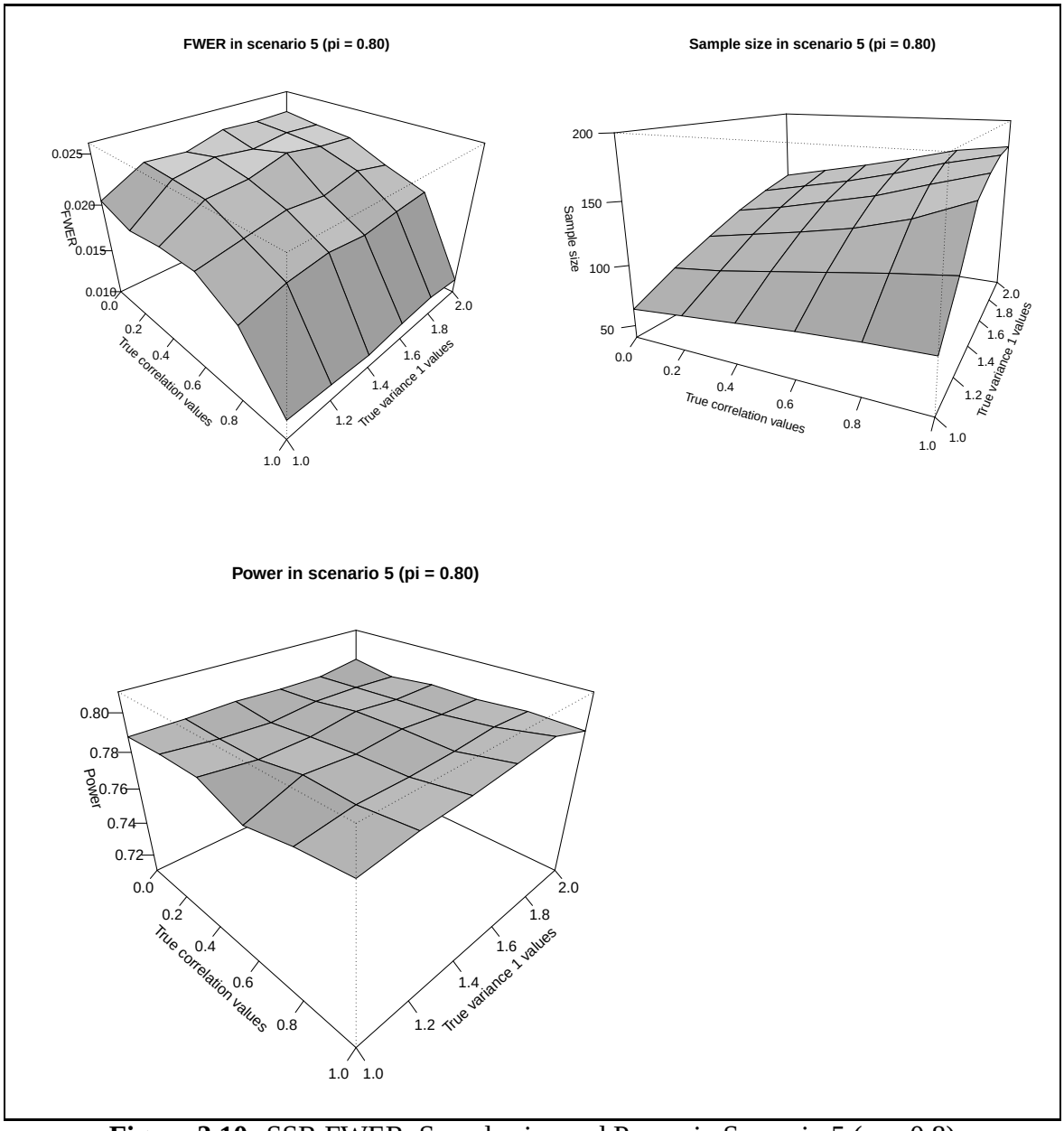

Figure 3.10: SSR FWER, Sample size and Power in Scenario 5 (π = 0.8)

conservative as ρ_{12} increases. The timings of the interim evaluation has no impact on the FWER. This has not been shown analytically although we have shown this by simulations in this scenario.

The results in Scenario 5 also show that sample sizes are increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2, and this is similar on both settings. However Figure 3.9 shows that the power is not maintained, while Figure 3.10 indicates that it is maintained. One of the explanations is given by Gould (1992), who stipulates that as the sample size at the interim review is decreasing, the likelihood of inadequate or unnecessarily large final samples sizes is increased. As can be seen in this scenario, although a similar sample size, Figure 3.9 gives a variable sample size, which depends on variable and unprecise nuisance parameters, so it is less powerful than the one in Figure 3.10 which gives a precise sample size based on precise nuisance parameters, hence powerful.

3.2 SSR Inverse Normal Combination test for multiple co-primary endpoints

This Section describes how the SSR inverse normal combination test method, described in Section 2.2, can be extended to the setting of K co-primary endpoints. It shows in more detail how to construct test statistics in such a way as to control the FWER if the variance σ_k^2 for endpoint *k* and the correlation $\rho_{kk'}$ between inverse normal tests is not known. Subsection 3.2.1 presents a framework of the analysis, Subsection 3.2.2 describes the formulation of problem considered, Subsection 3.2.3 presents test statistics, Subsection 3.2.4 illustrates the implementation of the method, Subsection 3.2.5 presents a worked example and Subsection 3.2.6 presents simulations results.

3.2.1 Framework for the analysis of the method

The framework of analysis developed in Subsection 2.2.3.1 is extended as follows in the context of multiple co-primary endpoints:

3.2.1.1 Step 1 - Initial sample size calculation

The initial sample size N_0 is calculated on the basis of an initial estimate of the nuisance parameters $\rho_{kk'_0}$ ($k' > k$) and $\sigma_{k_0}^2$ guessed before the trial begins. N_0 is computed as in Subsection 3.1.4.

3.2.1.2 Step 2 - Sample size re-estimation

When the data for the first $n_1 = \pi N_0$ patients (e.g., π = 0.5) are available, then the nuisance parameters $\rho_{kk'_1}$, $k' > k$ and $\sigma_{k_1}^2$ are re-estimated from these observations, which constitute the internal pilot study. These estimates are then used to calculate the sample size N. The new variance estimate $\sigma_{k_1}^2$ is used to calculate *t*-statistics T_{k1} for endpoint *k* calculate with

n_1 observations. T_{k1} is then used to calculate the *p*-value p_{k1} for endpoint *k*. A further $n_2 = N - n_1$ observations are collected and used to estimate the variance $\sigma_{k_2^2}$, which is used to calculate *t*-statistics T_{k2} for endpoint *k*. T_{k2} is then used to calculate *p*-value p_{k2} for endpoint *k*.

3.2.1.3 Step 3 - Final analysis

At the final analysis, the evidence from stage 1 and stage 2 are combined via the weighted inverse normal function B_k of the observed p_{k1} and p_{k2}. The hypothesis test is conducted using B_k; more details about this are given in the Subsection below.

3.2.2 Formulation of the problem

In this section, we take into account the same problem as in Subsection 3.1.2. We consider methodology for situations where there are K co-primary continuous correlated endpoints in a clinical trial. Suppose that E and C are two treatments to be compared in a randomised (phase III) parallel group clinical trial. After each group of *N* subjects has been randomised in equal numbers to the two therapies and the response obtained, $\rho_{kk'}$ and σ_k^2 are re-estimated and the accumulated data is tested. However, in contrast to the problem defined in Subsection 3.1.2, the accumulated data are now tested using a combination of inverse normal tests of p_{kj}-values, one of the p_{kj}-values being derived from the data collected at interim stage and the other $p_{k(j+1)}$ being derived from the independent new data collected after sample re-estimation. The primary trial's objective is to determine whether E is more efficacious than C in terms of K continuous co-primary responses. This procedure is conducted at the final step of SSR analysis framework described in Subsection 3.2.1.3, which involves a comparison of the evidence of efficacy of E and C, with the rejection occurring as soon as one of the K-hypotheses is in some sense sufficiently convincing.

3.2.3 Test statistics

In this Subsection, we are interested in constructing test statistics in the setting of K co-primary endpoints and deriving their distribution.

Suppose we use the test statistic B_k defined in Eq. (2.25), where w_j is a predefined weight and the *p*-value p_j is defined in Eq. (2.26). Let B_k, k = 1,..,K now denote the test statistic for θ_k and endpoint k and p_{kj} (j = 1,2) the *p*-value for endpoints *k* and stage *j*, which we write as:

$$p_{kj} = 1 - \Phi(Z_{kj}) \tag{3.14}$$

and

$$B_k = w_1\Phi^{-1}(1 - p_{k1}) + w_2\Phi^{-1}(1 - p_{k2}) \tag{3.15}$$

where Z_{kj} defined in Eq. (3.1) is now the standardised test statistics for endpoint *k* based on *new data* at stage *j*.

Under H_{0k}, B_k is normal distributed with mean 0 and variance $(w_1)^2 + (w_2)^2$, i.e. $B_k \sim N(0, (w_1)^2 + (w_2)^2)$.

Suppose $\rho_{kk'}$ is the correlation between endpoints :

$$Cov(B_k, B_{k'}) = \rho_{kk'}, k^{'} > k. \tag{3.16}$$

B_k (k = 1,. . ., K) has a multivariate normal distribution:

$$\begin{pmatrix} B_1 \\ \vdots \\ B_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} , \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ & 1 & \cdots & \rho_{2K} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right). \tag{3.17}$$

3.2.3.1 Implication for the FWER

We now need to show that using the critical value *c* defined in Eq. (1.15), the distribution of the inverse normal test statistics constructed under the null hypothesis in Eq. (3.17) controls the FWER in the strong sense.

So for one endpoint, in place of Eq. (2.31) we now define the type I error rate to be:

$$\begin{aligned} &Pr(\text{reject } H_0 \mid \theta = 0) = \\ &Pr(B_k \geq c) = \alpha/K. \end{aligned} \tag{3.18}$$

Here, the Bonferonni correction is applied as illustrated in Section 3.3.1.

In the setting of K endpoints, in place of Eq. (3.18), we now have:

$$\begin{aligned} &Pr(\text{reject at least one } H_{0k} \mid \theta_k = 0) = \\ &Pr(B_1 > c \text{ or,}\ldots,\text{or } B_K > c \mid \theta_k = 0) \leq \alpha. \end{aligned} \tag{3.19}$$

So, to control the FWER, one must use c to satisfy Eq. (3.19).

As explained previously, Whitehead (2010) stated that the specification and interpretation of alternative hypotheses is more difficult to define in general. However, this thesis consider using the sample size of the SSR method with multiple co-primary endpoints and

the SSR inverse normal combination test statistics with multiple co-primary endpoints to maintain the power. This is checked by simulations in Subsection 3.2.6.

3.2.4 Implementation of the method

This section illustrates how the framework of analysis described in Subsection 3.2.1 could be implemented in practice. It is a adaptation of the three-steps procedure developed in Subsection 3.1.1. The following is an illustration of such a method:

3.2.4.1 Step 1 - Initial sample size calculation

As in SSR with multiple co-primary endpoints setting, in this step, the initial sample size calculation leading to a provisional sample size N_0 is carried out on the basis of an initial estimate of the nuisance parameters.

B1.1. Guess $\rho_{kk_0'}$ ($k' > k$) and $\sigma^2_{k_0}$.

B1.2. Determine α and target power = 1 - β.

B1.3. Calculate the boundary c as in Eq. (1.15).

B1.4. Calculate the initial sample size N_0 as illustrated in more details in Subsection 3.1.4.

B1.5. Fix the fraction of the initial sample size π to be used in step 2.

B1.6. Pre-define the weight w_j (j= 1,2) in advance satisfying Eq. (2.20).

3.2.4.2 Step 2 - Sample size re-estimation

In this step, the interim data are used to re-estimated the nuisance parameters guessed at the design stage and determine the sample size for the remainder of the study. In this step, we also compute *p*-values, p_{k1} and p_{k2}, from the *t*-distribution.

B2.1. Simulate $n_1 = \pi N_0$ observations.

B2.2. Estimate $\sigma^2_{k_1}$ using blinded method as in Eq. (2.10), based on n_1 observations.

B2.3. Use $\sigma^2_{k_1}$ to calculate *t*-test T_{k1} for endpoint *k*.

B2.4. Calculate degrees of freedom df_1 as in Eq. (1.27), based on n_1 observations.

B2.5. Use T_{k1} and df_1 to calculate *p*-value p_{k1} as defined in Eq. (1.26).

B2.6. Estimate $\rho_{kk'_1}$ as in Eq. (1.9), based on n_1 observations.

B2.7. Use α and power defined in step (B1.2), *c* defined in step (B1.3), $\sigma^2_{k_1}$ and $\rho_{kk'_1}$ defined in steps (B2.2) and (B2.6) respectively to re-calculate sample size N as illustrated in Subsection 3.1.4.

B2.8. Collect a further $n_2 = N - n_1$ observations using a restricted design as described in Eq. (2.7).

B2.9. Estimate $\sigma^2_{k_2}$ using blinded method as in Eq. (2.10), based on n_2 observations.

B2.10. Use $\sigma^2_{k_2}$ to calculate *t*-test T_{k2} for endpoint *k*.

B2.11. Calculate degrees of freedom df_2 as in Eq. (1.27), based on new n_2 observations.

B2.12. Use T_{k2} and df_2 to calculate *p*-value p_{k2} as defined in Eq. (1.26).

3.2.4.3 Step 3 - Final analysis

At the final analysis, the evidence from stage 1 and stage 2 are combined via the weighted inverse normal function B_k of the observed p_{k1} and p_{k2}. The hypothesis test is conducted using test statistic B_k.

B2.1. Combine p_{k1} and p_{k2} by the weighted inverse normal function B_k

B2.2. Reject at lest one H_{0k} at level α if $B_k \geq c$.

3.2.5 Worked example of the method

In this subsection, the same example as in Subsection 3.1.6 is considered. Suppose that E and C are two treatments to be compared in a randomised (phase III) parallel group clinical trial. Two co-primary endpoints are considered, i.e. K = 2. Patients are randomised in equal numbers between E and C, and a normal distributed response is observed for each of the endpoints. Suppose the parameters of interest representing the mean differences are $\theta_1 = \theta_2 = 0.5$.

In step 1 (see Subsection 3.2.4.1 and Figure 3.1), we suppose (or guess) that the variance for endpoint 1 is $\sigma_{1_0} = 1.5$, the variance for endpoint 2 is $\sigma_{2_0} = 1$ and the correlation between endpoints is $\rho_{12_0} = 0.5$. A SSR inverse normal combination test with multiple co-primary endpoints is required to test $H_{0k} : \theta_k$ = 0, k = 1,2, with a one-sided test type I error rate of $\alpha = 0.025$ and a power of 1 - β = 0.80 for $\theta = 0.5$. We use step (B1.3) to calculate the critical value or boundary at level α/K. We then use step (B1.4) to estimate the initial sample size N_0. We finally fix the fraction of the initial sample size π = 0.5 as in step (B1.5) and the weight w_j (j= 1,2) = $\sqrt{0.5}$ as in step (B1.6).

In step 2, we simulate the interim data $n_1 = \pi N_0$ as illustrated in step (B2.1). Based on n_1 observations, the nuisance parameter $\sigma^2_{k_1}$ is re-estimated as in steps (B2.2), the *t*-statistic T_{k1} for endpoint *k* is calculated as in step (B2.3), the degrees of freedom df_1 is calculated as in step (B2.4) and *p*-value p_{k1} is calculated as in step (B2.5). The correlation $\rho_{kk'_1}$ is also re-estimated based on n_1 observations as in step (B1.6). The initial sample size N_0 is re-calculated by replacing the values of $\sigma^2_{k_1}$, ρ_{12_1} and c in the mean vector, the variance-covariance matrix and the upper limit matrix as illustrated in more detail in Subsection 3.1.6 and step (B2.7). This gives N. We continue by collecting a

Table 3.4: Scenarios considered in the simulation study.

Variable values	SSR inverse normal combination test
Scenario 6	**Different weights**
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Proportion π in SSR	0.50
True nuisance parameters	
ρ_{12}	0,0.1,...,1
σ_1^2	1,1.1,1.2,...,2
Setting 3	
σ_2^2	1.5
Weight for stage 1	0.1
Weight for stage 2	0.9

further $n_2 = N - n_1$ observations as in step (B2.8). Based on n_2 observations, $\sigma_{k_2}^2$ is estimated as in step (B2.9), *t*-statistic T_{k1} is calculated as in step (B2.10), the degrees of freedom df_2 is calculated as in step (B2.11) and *p*-value p_{k2} is calculated as in step (B2.12).

In step 3, we combine p_{k1} and p_{k2} by the weighted normal function B_k as in step (B3.1) and perform hypothesis testing by rejecting H_{01} or H_{02} at level α if $B_1 \geq c$ or $B_2 \geq c$ as in step (B2.2).

3.2.6 Simulation results

This subsection presents simulation results of the SSR inverse normal combination test method. On top of the scenarios considered in the Table 3.2, scenario 6 have been added to check the impact of different weights on FWER, sample size and power. This can be seen in Table 3.4.

3.2.6.1 FWER, power and sample size in the SSR inverse normal combination test design

The aim of this subsection is to check whether the FWER would be maintained and the power controlled if the nuisance parameters change as in the following settings:

3.2.6.1.1 Scenario 2 : FWER in Settings 1 - 5

Figure 3.11 presents FWER's simulation results with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. On top of that, an equal weight of 0.5 has been considered for stage 1 and stage 2 data. The figure shows that this method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 in all five settings. In setting 1, the FWER has a minimum value of 0.01036 for perfectly correlated data i.e. $\rho_{12} = 1$ and a maximum value of 0.02194 for uncorrelated data, i.e. $\rho_{12} = 0$. In setting 2, a minimum value of 0.01079 for $\rho_{12} = 1$ and a maximum value of 0.02292 for $\rho_{12} = 0$ have been observed. In setting 3, the FWER has a minimum value of 0.01046 for $\rho_{12} = 1$ and a maximum value of 0.02416 for $\rho_{12} = 0$. In setting 4, a minimum value of 0.01091 for $\rho_{12} = 1$ and a maximum value of 0.02378 for $\rho_{12} = 0$ have been observed. Finally, in setting 5, the FWER has a minimum value of 0.01066 for $\rho_{12} = 1$ and a maximum value of 0.02388 for $\rho_{12} = 0$.

3.2.6.1.2 Scenario 2 : Sample size in Settings 1 - 5

Figure 3.12 presents sample size simulation results with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. The results are presented in all five settings of scenario 2. The figure shows that the sample size increases as ρ_{12}, σ_1^2 and σ_2^2 increase. This is what we would expect. This also shows that the method is working.

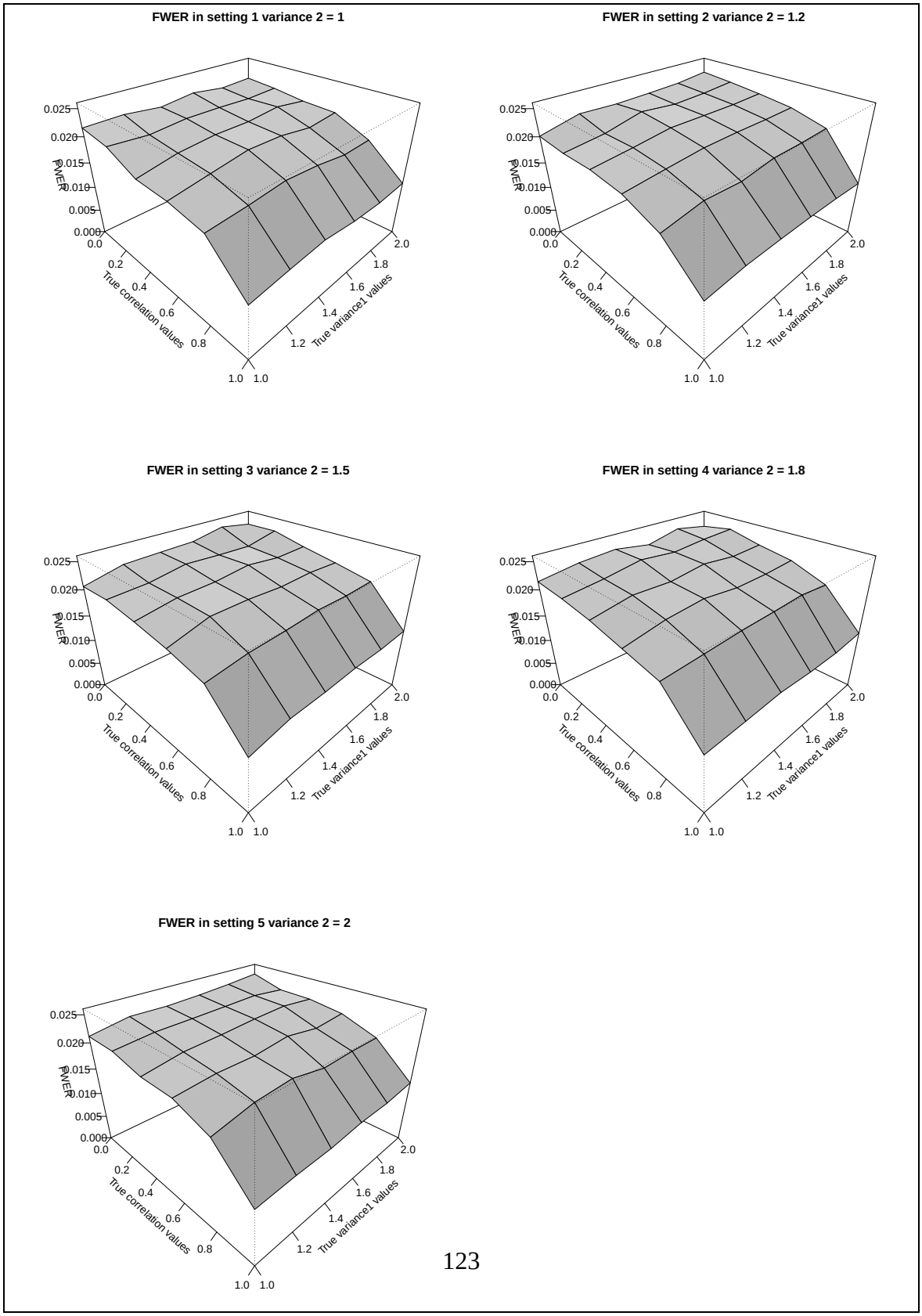

Figure 3.11: SSR Combination test FWER in Scenario 2; Settings 1 - 5

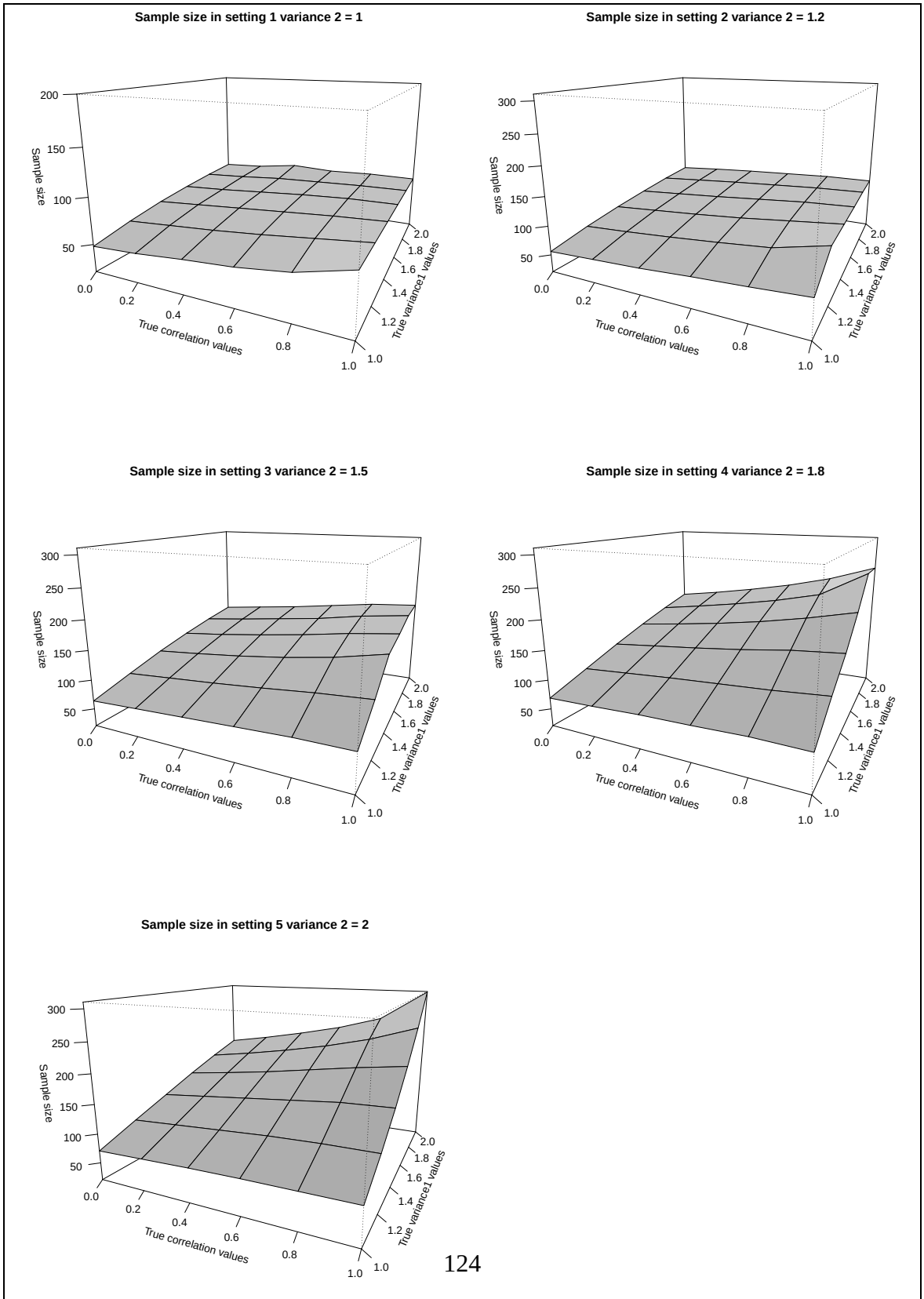

Figure 3.12: SSR Combination test Sample size in Scenario 2; Settings 1 - 5

3.2.6.1.3 Scenario 2 : Power in Settings 1 - 5

Figure 3.13 presents simulation results for the power with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively. The results are presented in all five settings of scenario 2. The figure illustrates that the method does not maintain the power. The figure shows that the power starts by increasing in setting 1, becomes constant in setting 2, then starts decreasing in settings 3 - 5 when ρ_{12}, σ_1^2 and σ_2^2 increase. This is because of inefficient weighting if SSR occurs.

3.2.6.1.4 Scenario 3: Constant ρ_{12}

The results in Scenario 3 are presented in Figure 3.14. It illustrates that despite variation of σ_1^2 and σ_2^2, the FWER is controlled and fairly constant with a minimum value of 0.01918 and a maximum value of 0.02260. The same figure illustrates that the sample size increases in the same direction as σ_1^2 and σ_2^2 with the minimum value 58 and the maximum value 220. Finally the figure illustrates that the power decreases when σ_1^2 and σ_2^2 increase with a minimum value of 0.7230 and a maximum value of 0.7810.

3.2.6.1.5 Scenarios 2 and 3: Summary and comments on the results

The results in Scenario 2 show that the FWER is controlled, however they show that the FWER becomes increasingly conservative as ρ_{12} increases. The results also show that the FWER is controlled and fairly constant when ρ_{12} is constant (Scenario 3). This is not surprising as it has been shown that the combination test method maintains the type I error rate for any possibly data-driven choice of sample sizes as illustrated in more detail in Subsection 2.2.4.2.

The results in Scenarios 2 and 3 also show that the sample size is increasing in the

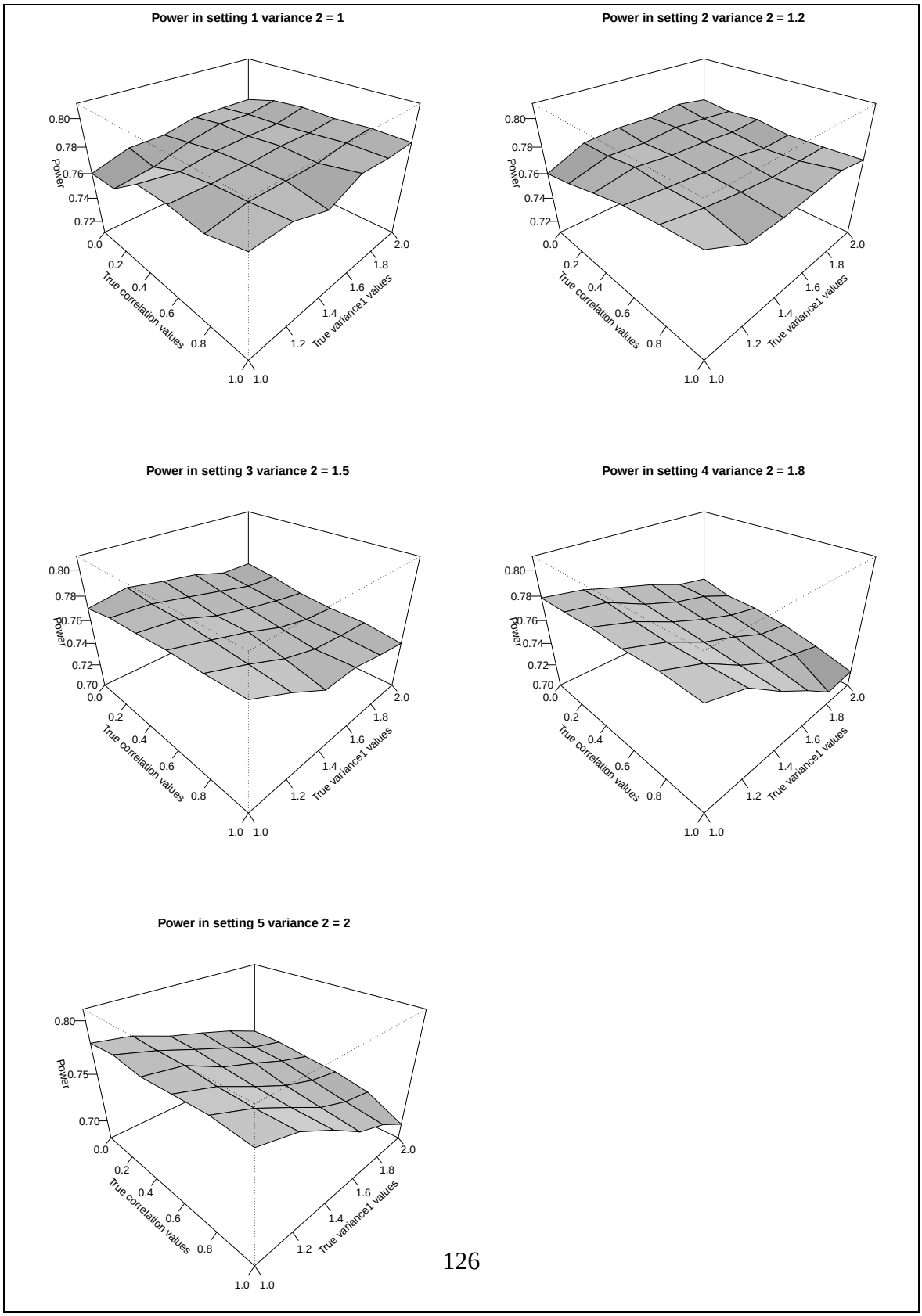

Figure 3.13: SSR Combination test Power in Scenario 2; Settings 1 - 5

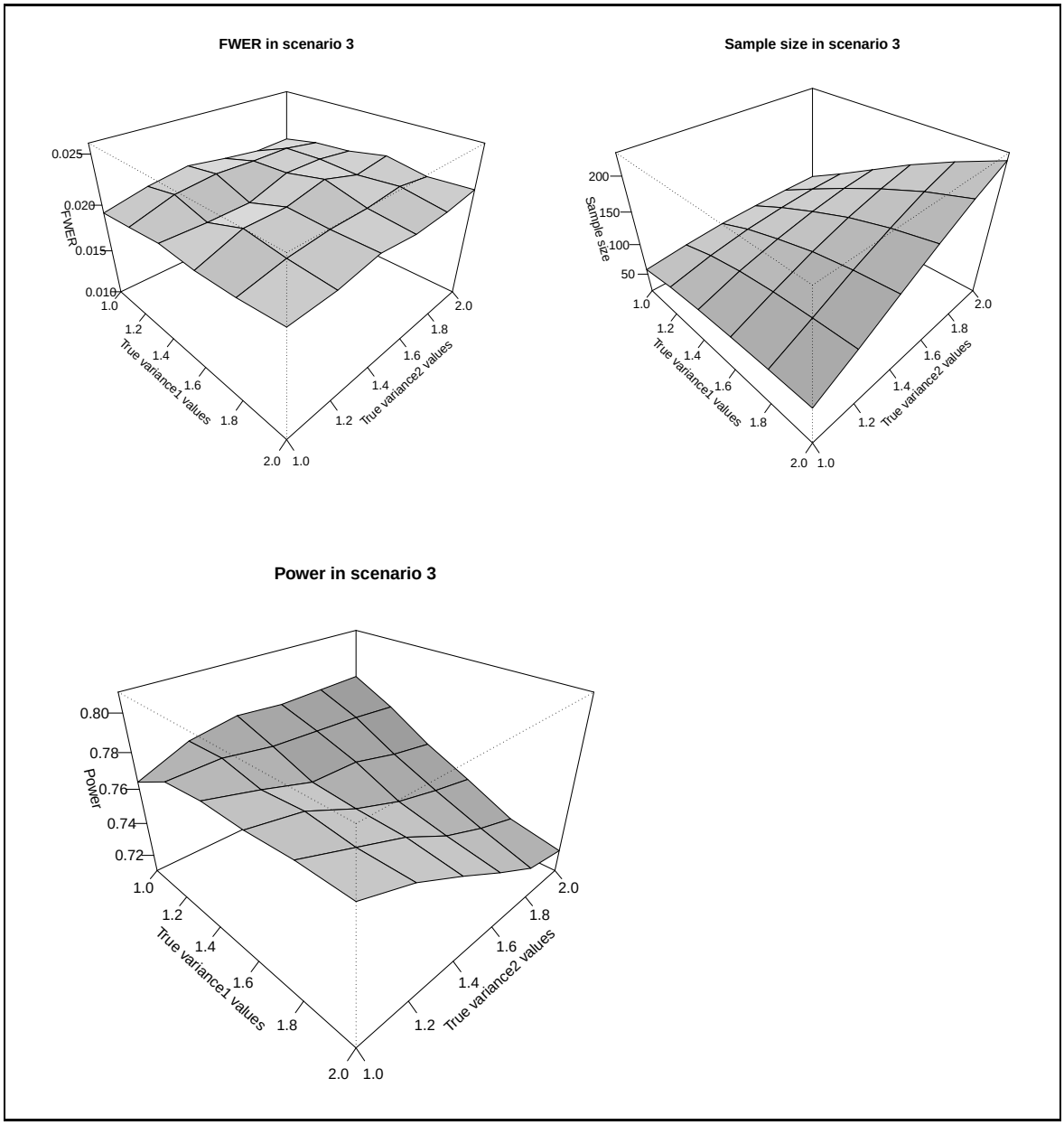

Figure 3.14: SSR Combination test FWER, Sample size and Power in Scenario 3

same direction as ρ_{12}, σ_1^2 and σ_2^2, however, the power is not maintained despite the fact that this method is adjusting for the sample size needed to maintain the power at least up to the nominal level. The reason for this is that the method use pre-defined weights which are not based on observed sample size but on weights fixed in advance. The paper by Bank et al. (1996) has shown this.

3.2.6.2 Scenario 4: Difference effect sizes

3.2.6.2.1 Scenario 4: $\delta_1 = 0.5, \delta_2 = 0.7$

Figure 3.15 presents simulation results with the fixed and variable values defined in Table 3.2 and Table 3.3 respectively; and a weight of 0.5 for stage 1 and stage 2 data. The situation of $\delta_1 = 0.5, \delta_2 = 0.7$ is considered. The figure shows that the FWER is controlled despite variation of ρ_{12} and σ_1^2 with the minimum value 0.01024 for perfect correlated data and 0.02151 for uncorrelated data. The same figure shows that the sample size increases as ρ_{12} and σ_1^2 increase with a minimum value of 51 and maximum value of 94. Finally the figure shows that the power is not maintained but fairly constant when ρ_{12} and σ_1^2 increase with a minimum value of 0.7270 and a maximum value of 0.7490.

3.2.6.2.2 Scenario 4: $\delta_1 = 0.7, \delta_2 = 0.5$

Figure 3.16 presents the situation where $\delta_1 = 0.7$ and $\delta_2 = 0.5$ are considered. The figure shows that the FWER is controlled despite variation of ρ_{12} and σ_1^2 with the minimum value 0.00788 for $\rho_{12} = 1$ and 0.02318 for $\rho_{12} = 0$. The same figure shows that the sample size increases as ρ_{12} and σ_1^2 increase with a minimum value of 39 and maximum value of 160. Finally the figure shows that the power is not maintained. It starts by increasing up to the peak for the value of $\sigma_1^2 = 1.2$ then decreases afterwards with a minimum value of 0.6734 and a maximum value of 0.7784.

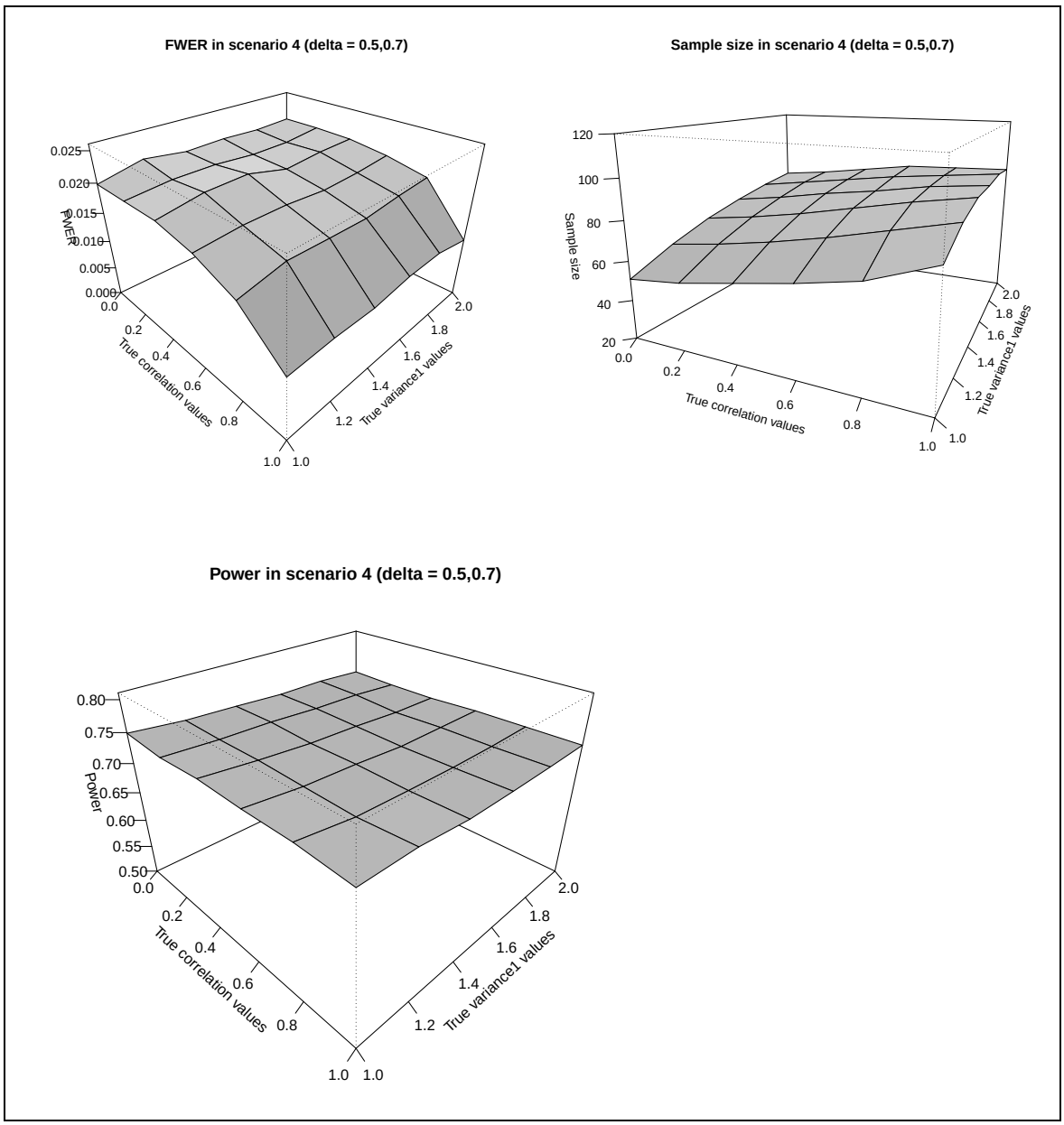

Figure 3.15: SSR Combination test FWER, Sample size and Power in Scenario 4 ($\delta_1 = 0.5$, $\delta_2 = 0.7$)

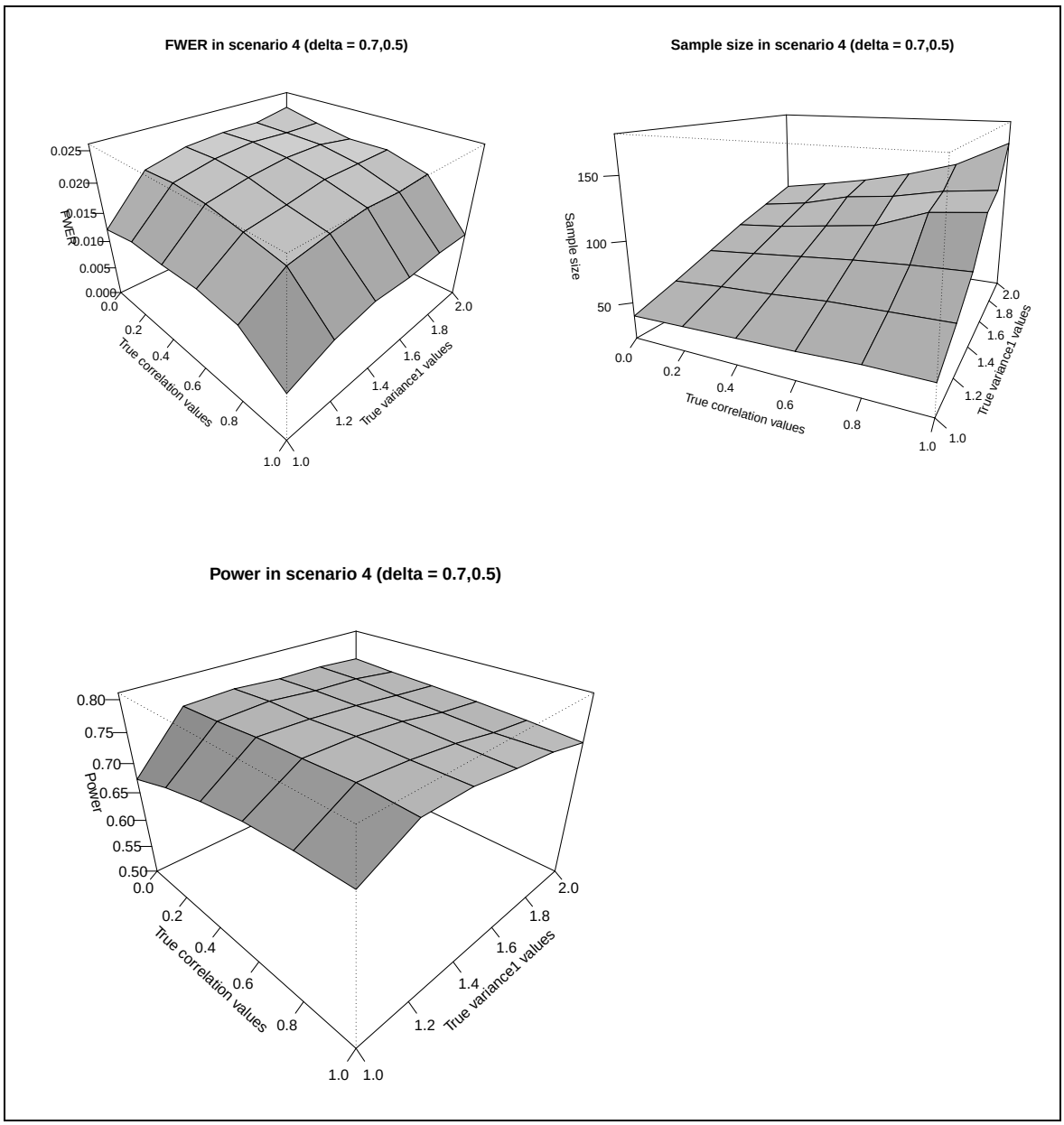

Figure 3.16: SSR Combination test FWER, Sample size and Power in Scenario 4 ($\delta_1 = 0.7$, $\delta_2 = 0.5$)

3.2.6.2.3 Scenario 4: Summary and comments on the results

In scenario 4 the results show that the FWER is controlled but conservative as ρ_{12} increases.

The results in scenario 4 also show that sample sizes are increasing in the same direction as ρ_{12} and σ_1^2, with setting (0.7,0.5) giving a larger sample size than the setting (0.5,0.7). However, the power in both settings is not maintained due to the use of pre-defined weights fixed before the trial begins and for the same reasons described in Subsection 3.1.7.3.3.

3.2.6.3 Scenario 5: Different timings

3.2.6.3.1 Scenario 5: $\pi = 0.10$

Figure 3.17 presents the simulation results for a situation where the time of interim evaluation happens when 10% of the data have been collected .i.e, $\pi = 0.10$. An equal weight of 0.5 for the two stages is also considered. The figure shows that this method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 with the minimum value 0.00932 for perfectly correlated data and the maximum value 0.01828 for uncorrelated data. The same figure also shows that the sample size increases as ρ_{12} and σ_1^2 increase with a minimum value of 61 and maximum value of 176. Finally the figure shows that the method does not maintain the power although this is fairly constant despite variation of ρ_{12} and σ_1^2 with a minimum value of 0.5850 and maximum value of 0.6406.

3.2.6.3.2 Scenario 5: $\pi = 0.80$

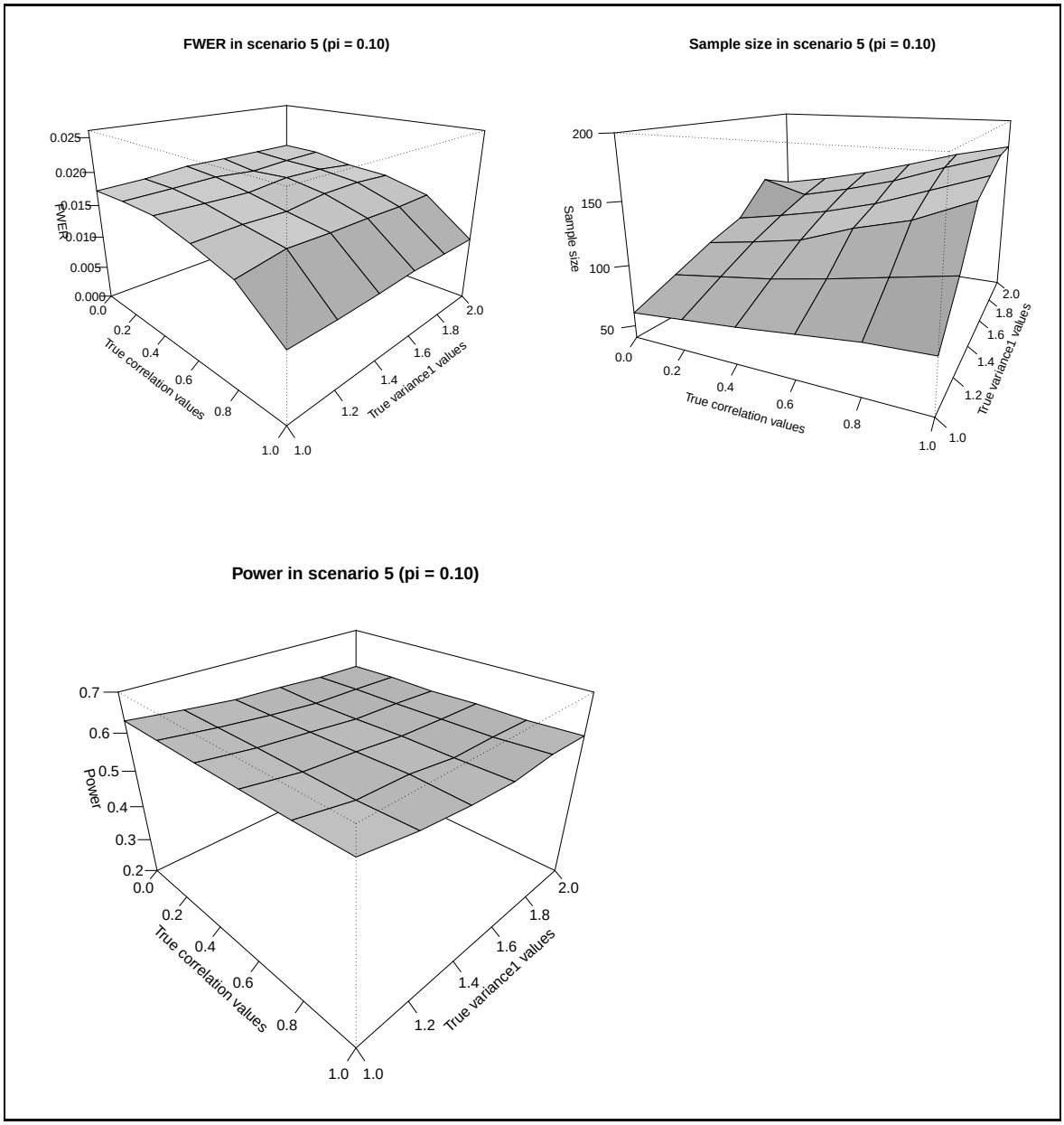

Figure 3.17: SSR Combination test FWER, Sample size and Power in Scenario 5 (π = 0.1)

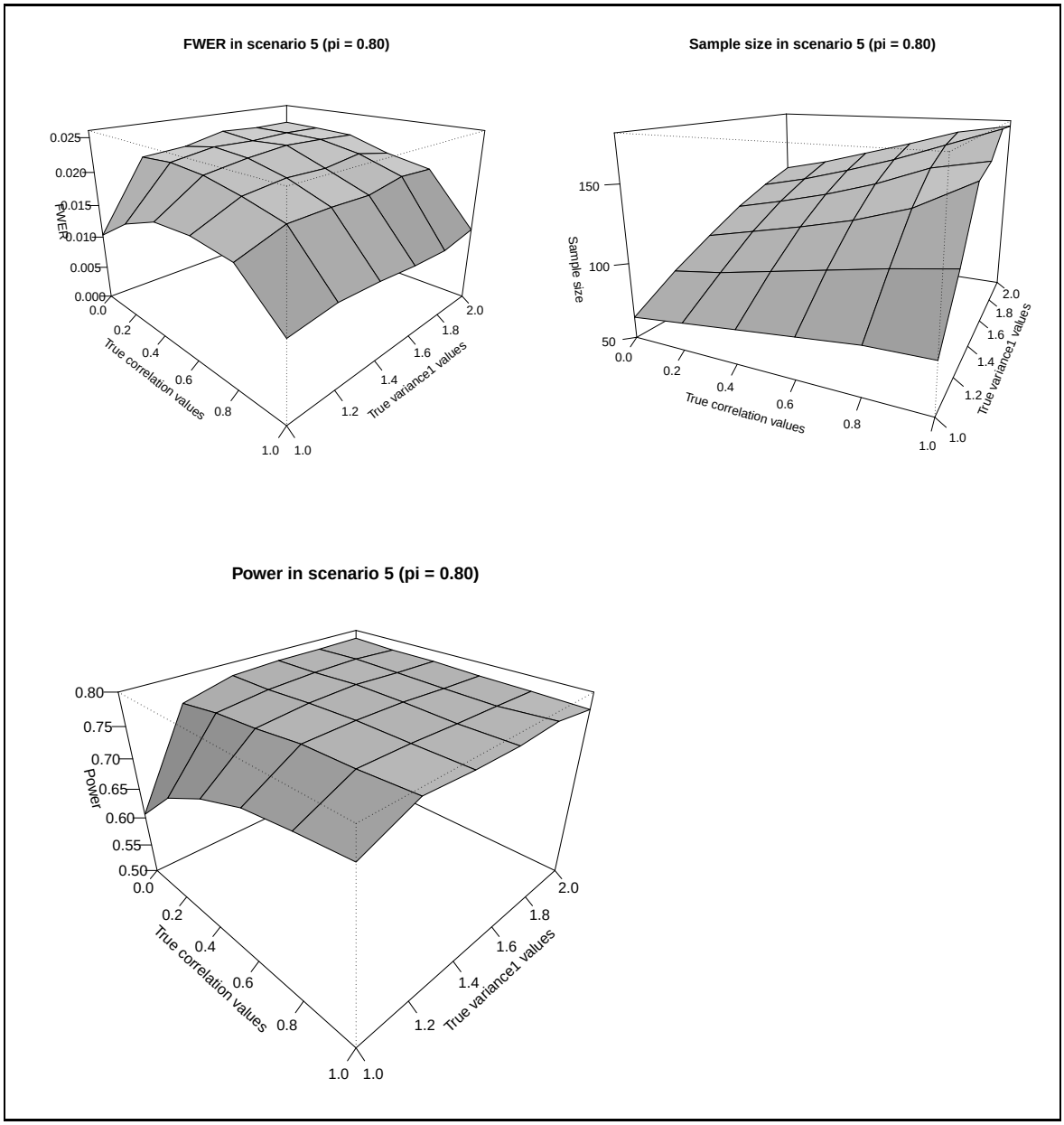

Figure 3.18: SSR Combination test FWER, Sample size and Power in Scenario 5 (π = 0.8)

In Figure 3.18, a situation where the time of interim evaluation happens is set to $\pi = 0.80$ and an equal weight of 0.5 for the two stages is also considered. The figure shows that this method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 with the minimum value 0.01021 for $\rho_{12} = 1$ and the maximum value 0.02300 for $\rho_{12} = 0$. The same figure also shows that the sample size increases as ρ_{12} and σ_1^2 increase with a minimum value of 64 and maximum value of 177. Finally the figure shows that the method does not maintain the power for the combination of $\rho_{12} = (0, 0.1, ..., 1)$ and $\sigma_1^2 = 1$ but does maintain it for the remaining of the combination of ρ_{12} and σ_1^2 with a minimum value of 0.6063 and a maximum value of 0.7894.

3.2.6.3.3 Scenario 5: Summary and comments on the results

The results in Scenario 5 show that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases. The timing of the interim evaluation has no impact on the FWER.

The same results show that sample sizes are increasing in the same direction as ρ_{12} and σ_1^2. However Figure 3.17 shows that the power is not maintained, while Figure 3.18 indicates that it is not maintained only for the value of $\sigma_1^2 = 1$ and maintained for any other values. Although we would expect this method not to maintain the power because it uses pre-defined weights which are not based on observed sample size but on weights fixed in advance, the results in Figure 3.18 shows that by changing the timing of the interim intervention, this method maintains the power even if it uses pre-defined weights fixed in advance.

3.2.6.4 Scenario 6: Different weights

We have further simulations to check the effect of unequal weights on FWER, sample size and power. Figure 3.19 presents simulation results with the fixed and variable values

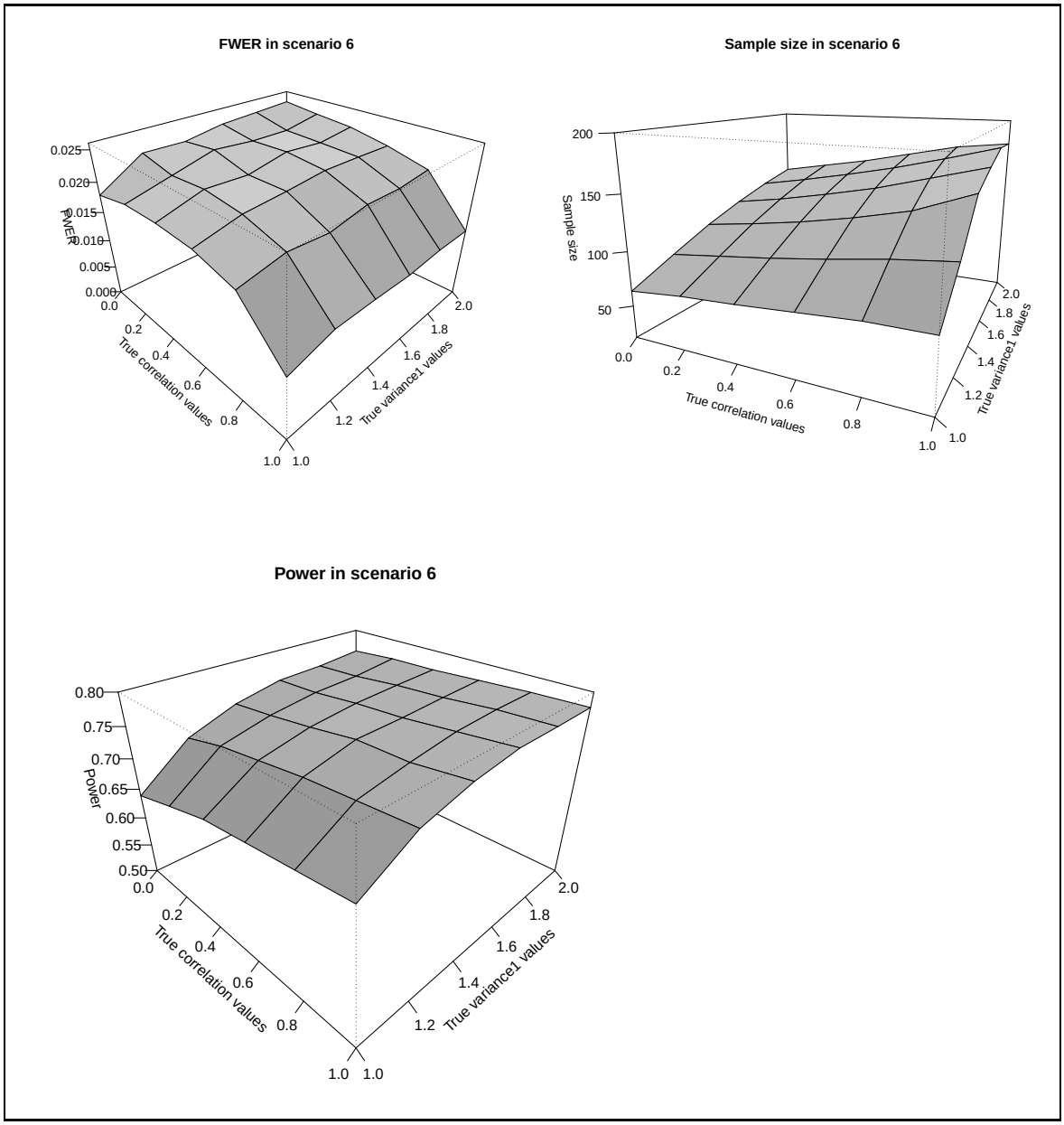

Figure 3.19: SSR Combination test FWER, Sample size and Power in Scenario 6

defined in Table 3.2 and Table 3.4 respectively; and a weight of 0.1 for stage 1 data and 0.9 for stage 2 data. The figure shows that despite variation of ρ_{12} and σ_1^2; and unequal weights allocation, the FWER is controlled with a minimum value of 0.01027 for $\rho_{12} = 1$ and a maximum value of 0.02408 for $\rho_{12} = 0$. The same figure shows that the sample size increases in the same direction as ρ_{12} and σ_1^2 with the minimum value 64 and the maximum value 178. Finally the figure illustrates that the power is not maintained but increases when ρ_{12} and σ_1^2 increase with a minimum value of 0.639 and a maximum value of 0.779.

3.2.6.4.1 Scenario 6: Summary and comments on the results

The results in this scenario (see Figure 3.19) are compared to the ones in scenario 2, setting 3. The results in both settings show that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases. Equal or unequal allocation of the pre-defined weights have no impact on the FWER.

The same results show that sample sizes are increasing in the same direction as ρ_{12} and σ_1^2. However Figure 3.13, setting 3 shows that the power is not maintained and this decreases when ρ_{12}, σ_1^2 and σ_2^2 increase, while Figure 3.19 show that the power is not maintained but increases when ρ_{12} and σ_1^2 increase. The results in Figure 3.19 show that changing the weights allocation has an impact on maintaining the power, however, finding the optimal weights allocation is challenging because they are defined before the trial begins.

3.3 Summary findings from the simulation results

This chapter presented the idea of sample size reestimation in the context of multiple co-primary endpoints. In Section 3.1, we presented a sample size reestimation approach for this setting and in Section 3.2 we described the inverse normal combination tests method

with multiple co-primary endpoints with sample size reestimation. Simulation results have been presented in Sections 3.1.7 and 3.2.6 respectively. This section summarize key findings from these results.

In Section 3.1.7, we observed that the SSR with multiple co-primary endpoints method controls the FWER; and this is true for all the scenarios considered. For example, in Scenario 2 we noticed no evidence of an inflation of the FWER for all the settings studied. We noticed similar results in Scenarios 3,4 and 5. These results are comparable to the results obtained by Friede and Schmidli (2010). We also observed that, although the FWER was controlled in Scenarios 2,4 and 5, it (FWER) became increasingly conservative as ρ_{12} increased. These results are similar to the findings obtained by Senn and Bretz (2007). Therefore we concluded that for the settings considered here, the SSR procedure controls the FWER.

In Section 3.1.7, we also observed that the sample size was increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2.

In Section 3.1.7, we finally observed that in most scenarios, the method maintains the power. However, we observed that if strange revision rules are used, the power could not be maintained. For example, we noticed in Scenario 4 that the SSR method does not maintain the power when different effect sizes are used simultaneously. The key findings for this scenario were that the magnitude of the sample size for the SSR method was driven by the big effect size; and this (sample size) was not large enough to detect two different effect sizes at the same time, hence reduction in power. We recommended to use the small effect size for sample size calculation which is a guarantee that the sample size obtained would be large enough to detect even the large effect size and maintain the power. We also noticed that in Scenario 5 - Figure 3.9, the interim evaluation of the nuisance parameters was performed when 10% of the observations were in the internal pilot. This led to inac-

curate estimation of the nuisance parameters, with the consequence that the power was not maintained. This finding was investigated by Friede and Schmidli (2010) who explained that the timing of the interim evaluation in clinical trials is not only important for logistic motivations but also affects the operational characteristics of the recalculation procedure. An early interim review cannot give a good estimate of the nuisance parameters, while a very late sample size review may lead to a larger sample size than needed. The finding was also investigated by Gould (1992), who stipulated that as the sample size at the interim review is decreasing, the likelihood of inadequate or unnecessarily large final samples sizes is increased. As was seen in Scenario 5, although a similar sample size in both settings, Figure 3.9 gave a variable sample size, which depends on variable and unprecise nuisance parameters, so it was less powerful than the one in Figure 3.10 which gave a precise sample size based on precise nuisance parameters, hence powerful.

In Section 3.2.6, we observed that the SSR inverse normal combination test method controls the FWER; and this is true for all the scenarios considered. Although we have shows this by simulations, the papers by Bauer (1989), Lehmacher and Wassmer (1999) and Proschan (2009b) have proven analytically that the combination test method controls the type I error rate for any possibly data-driven choice of sample sizes.

In Section 3.2.6, we also observed that the sample size was increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2. In Section 3.2.6, we finally observed that the although the method controls the FWER, there is a cost, which is a loss of power, because the weights of the combination test are not based on the observed sample size but on weights fixed in advance. The papers by Lehmacher and Wassmer (1999), Proschan (2009b) and Bank et al. (1996) have shown this. We have noticed that although we would expect this method not to maintain the power for the reasons explained earlier, simulation results in Figure 3.18 showed that by changing the timing of the interim intervention, this method maintained the power even if it uses pre-defined weights fixed in advance. We finally noticed in Scenario

6 that changing the weights allocation had an impact on maintaining the power, however, finding the optimal weights allocation is challenging because they (weights) are defined before the trial begins.

As a conclusion, we showed that the results presented in this chapter clearly indicated that the FWER for both methods was controlled, regardless of the scenarios used. However, the results showed that in most scenarios considered, the SSR method was more powerful compared to the SSR inverse normal combination test method. In the next chapter, we are going to present how the group sequential method with a single endpoint described in Section 2.3 can be extended to the setting of multiple co-primary endpoints.

Chapter 4

Group Sequential Designs with Multiple Co-primary Endpoints

This chapter describes how the group sequential method with a single endpoint described in Section 2.3 can be extended to the setting of multiple co-primary endpoints. In other words, this chapter shows how a sequential clinical trial can be conducted in the presence of multiple co-primary endpoints. In this thesis we propose one way of doing this, which is to conduct *K* different group sequential tests, each for one endpoint and each adjusted at level α/K, while monitoring the information. The information is adjusted to allow for $\rho_{kk'}$ and σ_k^2 to be estimated at each stage (i.e. we know that at the beginning of the trial the endpoints are correlated with unknown $\rho_{kk'}$ and σ_k^2). After an introduction in Section 4.1, Section 4.2 describes GSD methodology for multiple endpoints, where the problem defined in Subsection 1.4.2 is considered. Section 4.3 presents the implementation of the method followed a summary in Section 4.4. Section 4.5 presents a worked example and Section 4.6 simulation results.

4.1 Introduction

In numerous phase III clinical trials, it is appropriate to distinctly assess the treatment effect on two or more primary endpoints. Kosorok et al. (2004) quoted an example of the MERIT-HF study, where two endpoints of primary interest were time to death and the earliest of time to first hospitalisation or death (on Behalf of the MERIT-HF Study Group (1997)). It is possible that treatment has no effect on death but a beneficial effect on first hospitalisation time, or it has a detrimental effect on death but no effect on hospitalization. A good clinical trial should permit early stopping as soon as the treatment effect on either endpoint becomes clear.

Depending on the nature of the endpoints, there are two statistical methods for multiple outcomes in group sequential clinical trials: (i) *global methods* that attempt to combine several endpoints into a single endpoint or statistical test; and (ii) *multiple hypothesis methods* that allow the assessment of differential treatment effects in two or more outcomes.

4.1.1 Global methods

If a similar aspect of treatment performance is measured by several endpoints, then demonstrating efficacy in a collective way across the endpoints would be acceptable and *global method* could be used. Examples and applications of this procedure include O'Brien (1984), Wei and Lachin (1984), Pocock, Geller and Tsiatis (1987), Tang,Gnecco and Geller (1989), Lin (1991), and Block, Lai, and Tubert-Bitter (2001). O'Brien (1984) describes a randomised clinical trial of two therapies for the treatment of diabetes patients in which improvements in nerve function were measured on 34 electromyographic variables combined as a single variable. Pocock, Geller and Tsiatis (1987) consider a crossover trial of chronic respiratory disease in which the active drug and placebo were compared with respect to three lung function measurements: (peak expiratory flow rate, forced expiratory

volume and forced vital capacity). This is a case where the outcome variables might be regarded as unordered since one does not take priority over another. Another example is the standard approach to multivariate outcomes in clinical trials developed by Pocock (1997) which includes the method of constructing a single composite endpoint from two or more outcomes.

4.1.2 Multiple hypothesis methods

With some endpoints, it is not clinically meaningful to combine the endpoints into a single measure; in this case, multiple hypothesis methods would be appropriate (Kosorok et al. (2004)). Stallard and Todd (2010) give an example of Alzheimer's disease where the need to demonstrate efficacy for two outcomes to gain licensing requires the consideration of both mental and physical aspects of a patient's progress. They also give an example of a simultaneous monitoring of efficacy and safety responses where the endpoints are kept separate and the correlation between them are accounted for in the design and analysis of any sequential trial.

Some other examples and applications include Jennison and Turnbull (1993), who proposed group sequential tests for bivariate response. Their tests are defined in terms of the two response components jointly, rather than through a single summary statistic. Such methods are appropriate when the two responses concern different aspects of a treatment; for example, one might wish to show that a new treatment is both as effective and safe as the current standard. They present a formulation of the bivariate testing problem, introduce group sequential tests that satisfy type I error conditions and show how to find the sample size ensuring a specified power. Jennison and Turnbull's (1993) approach is similar to what we are proposing as they consider the case of two primary endpoints, which are efficacy and safety outcomes. However, in terms of the types of analysis, Jennison and Turnbull's

(1993) method is different in the way that our method allows early stopping as soon as the treatment effect on *either* endpoint becomes clear, while theirs permit early stopping as soon as the treatment effect on *both* endpoints becomes obvious.

Cook and Farewell (1994) propose methodological guidelines for the sequential assessment of experimental therapies formally based on both efficacy and toxicity responses. It is based on a modified univariate sequential procedure (e.g., Lan and DeMets (1983)) accounting for bivariate correlated responses. It differs from Jennison and Turnbull's (1993) approach in that it is a rejective test and does not involve specification of an indifference region in the efficacy-toxicity parameter space. In terms of types of multiple endpoints, both methods are similar to what we are proposing as they consider the case of two primary endpoints, which are efficacy and safety outcomes for Jennison and Turnbull (1993) and efficacy and toxicity endpoints for Cook and Farewell (1994). However, in terms of the types of analysis, both methods are different to what we are proposing in the way that our method allows early stopping as soon as the treatment effect on either endpoint becomes clear, while theirs permit early stopping as soon as the treatment effect on both endpoints becomes obvious.

Glimm et al. (2010) propose a method of testing hierarchically a (key) secondary endpoint in a group-sequential clinical trial that is mainly driven by a primary endpoint. By mainly driven, they mean that the interim analyses are planned at points in time where a certain number of patients or events have accrued on the primary endpoint, and the trial will run either until statistical significance of the primary endpoint is achieved at one of the interim analyses or to the final analysis. They also consider the situation where the trial is stopped as soon as the primary endpoint is significant, as well as the situation where it is continued beyond primary endpoint significance to further investigate the secondary endpoint. In addition, they investigate how to achieve strong control of the familywise error rate (FWER) at a pre-specified significance level α for both primary and secondary hy-

potheses. A similar problem is proposed by Tamhane et al. (2010), who suggest a method of testing hierarchically a primary and secondary endpoint in clinical trial where the secondary endpoint is tested only if the primary endpoint is significant. The trial uses a group sequential procedure with two stages. Glimm et al. (2010) address the same problem and reach similar conclusions, but do not give explicit analytical results concerning the FWER and the primary and secondary critical boundaries as the setting in Tamhane et al. (2010). These two methods are completely independent and complementary to each other. In terms of types of endpoint and types of analysis, both examples are different to what we are proposing such that their outcome variables might be regarded as ordered as one does take priority over the other, while our outcomes are considered as equal.

Other multiple hypothesis approaches include Todd (1987) and Conaway and Petroni (1995). The application to survival data was developed by Cook (1994) and Williams (1996). Tang and Geller (1999) proposed a closed testing procedure for multiple comparisons that allows the evaluation of significance of each endpoint individually. More recent works are presented by Tamhane et al. (2012a) and Tamhane et al. (2012b).

4.2 Group Sequential Designs with multiple co-primary endpoints

As explained in the introduction, the aim of this thesis is to answer two questions: First, how to adjust a sample size in a clinical trial with multiple continuous primary endpoints using adaptive and group sequential designs. Second, how to construct a test in such a way to control the FWER and maintain the power, even if the correlation ρ between endpoints is not known. The following method is proposed to resolve this: K different group sequential tests are conducted, each for one endpoint and each at level α/K, and they are conducted to

monitor the information and also where the information is adjusted to allow for estimation of correlation ρ at each stage.

In the following section, we start by formulating a multivariate testing problem, followed by the construction of group sequential tests that satisfy FWER conditions, then show how to find the sample size that guarantees a specified power.

4.2.1 Definition of the problems

In this Section, we consider methodology for situations where there are K co-primary correlated endpoints in a clinical trial. The general setting for this problem is defined in Section 3.5. Suppose that E and C are two treatments to be compared in a randomised (phase III) clinical trial with parallel groups. After each group of 2*n* subjects has been randomised in equal numbers to the two therapies and the response obtained, the nuisance parameters $\rho_{kk'}$ and σ_k^2 are re-estimated, the sample size re-estimated, the boundaries adjusted based on the re-estimated sample size and the accumulated data tested. The trial's primary objective is to determine whether E is more efficacious than C in terms of K continuous responses. This procedure is conducted at a sequence of up to J interim analyses, each involving a comparison of the evidence for efficacy of E and C, with stopping occurring as soon as one of the interim analyses is in some sense sufficiently convincing.

4.2.2 Test statistics

Suppose that we are interested in repeated looks at the accumulating data on the co-primary endpoints with repeated hypothesis testing. At each interim analysis we will base inference on some calculated test statistics. We need to think about these test statistics and their distributions.

In Subsection 2.3.2.2, assumptions underlying many group sequential methods are presented in Eq. (2.38). In this subsection, we define test statistics for the setting of K endpoints and J stages, derive distributions and show how this relates to the canonical form.

We use the standardized statistic defined in Eq. (3.1) and the information for θ defined in Eq. (3.2) to construct our test statistics, and we assume σ_k^2 is known. Let Z_{kj}, k = 1,..,K and j = 1,...,J now denote the standardised statistic for θ_k and endpoint k based on the data available at interim analysis *j*, which we write as:

$$\begin{aligned} Z_{kj} &= \frac{1}{\sqrt{(2n_j\sigma_k^2)}}(\sum_i^{n_j} X_{ijkE} - \sum_i^{n_j} X_{ijkC}) \\ &\sim N((\theta_{kE} - \theta_{kC})\sqrt{\{n_j/(2\sigma_k^2)\}}, 1) \end{aligned} \tag{4.1}$$

under H_{0k}, Z_{kj} is normally distributed with mean 0 and variance 1, i.e., $Z_{kj} \sim N(0,1)$.

Let I_j now denote the information for θ_k based on the data available at interim analysis *j*, that is:

$$I_j = \frac{n_j}{2\sigma_k^2} \tag{4.2}$$

Suppose the correlation between test statistics is:

$$Cov(Z_{kj}, Z_{k'j}) = \rho_{kk'}, k' > k \tag{4.3}$$

and:

$$Cov(Z_{kj}, Z_{k'j'}) = \rho_{kk'}\sqrt{\frac{I_j}{I_{j'}}}, k' > k, j' > j. \tag{4.4}$$

As for Eq. (2.38), $Z_{11}, \ldots, Z_{K1}, \ldots, Z_{1J}, \ldots, Z_{KJ}$ has a multivariate normal distribution, at least asymptotically:

$$\begin{pmatrix} Z_{11} \\ \vdots \\ Z_{K1} \\ Z_{12} \\ \vdots \\ Z_{K2} \\ \vdots \\ Z_{1J} \\ \vdots \\ Z_{KJ} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1\sqrt{I_1} \\ \vdots \\ \theta_K\sqrt{I_1} \\ \theta_1\sqrt{I_2} \\ \vdots \\ \theta_K\sqrt{I_2} \\ \vdots \\ \theta_1\sqrt{I_J} \\ \vdots \\ \theta_K\sqrt{I_J} \end{pmatrix} , \left(\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ & 1 & \cdots & \rho_{2k} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \bigotimes \begin{pmatrix} 1 & \sqrt{\frac{I_1}{I_2}} & \cdots & \sqrt{\frac{I_1}{I_J}} \\ & 1 & \cdots & \sqrt{\frac{I_2}{I_J}} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right) \right) \tag{4.5}$$

In the case of K = 2 and J = 3, Eq. (4.5) can be expressed as:

$$\begin{pmatrix} Z_{11} \\ Z_{21} \\ Z_{12} \\ Z_{22} \\ Z_{13} \\ Z_{23} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1\sqrt{I_1} \\ \theta_2\sqrt{I_1} \\ \theta_1\sqrt{I_2} \\ \theta_2\sqrt{I_2} \\ \theta_1\sqrt{I_3} \\ \theta_2\sqrt{I_3} \end{pmatrix} , \begin{pmatrix} 1 & \rho & \sqrt{\frac{I_1}{I_2}} & \rho\sqrt{\frac{I_1}{I_2}} & \sqrt{\frac{I_1}{I_3}} & \rho\sqrt{\frac{I_1}{I_3}} \\ \rho & 1 & \rho\sqrt{\frac{I_1}{I_2}} & \sqrt{\frac{I_1}{I_2}} & \rho\sqrt{\frac{I_1}{I_3}} & \sqrt{\frac{I_1}{I_3}} \\ \sqrt{\frac{I_1}{I_2}} & \rho\sqrt{\frac{I_1}{I_2}} & 1 & \rho & \sqrt{\frac{I_2}{I_3}} & \rho\sqrt{\frac{I_2}{I_3}} \\ \rho\sqrt{\frac{I_1}{I_2}} & \sqrt{\frac{I_1}{I_2}} & \rho & 1 & \rho\sqrt{\frac{I_2}{I_3}} & \sqrt{\frac{I_2}{I_3}} \\ \sqrt{\frac{I_1}{I_3}} & \rho\sqrt{\frac{I_1}{I_2}} & \sqrt{\frac{I_2}{I_3}} & \rho\sqrt{\frac{I_2}{I_3}} & 1 & \rho \\ \rho\sqrt{\frac{I_1}{I_3}} & \sqrt{\frac{I_1}{I_3}} & \rho\sqrt{\frac{I_2}{I_3}} & \sqrt{\frac{I_2}{I_3}} & \rho & 1 \end{pmatrix} \right) \tag{4.6}$$

Replacing Eq. (4.2) into equation Eq. (4.5) and Eq. (4.6) respectively, we have:

$$\begin{pmatrix} Z_{11} \\ \vdots \\ Z_{K1} \\ Z_{12} \\ \vdots \\ Z_{K2} \\ \vdots \\ Z_{1J} \\ \vdots \\ Z_{KJ} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1 \sqrt{\frac{n_1}{2\sigma_1^2}} \\ \vdots \\ \theta_K \sqrt{\frac{n_1}{2\sigma_K^2}} \\ \theta_1 \sqrt{\frac{n_2}{2\sigma_2^2}} \\ \vdots \\ \theta_K \sqrt{\frac{n_2}{2\sigma_K^2}} \\ \vdots \\ \theta_1 \sqrt{\frac{n_J}{2\sigma_1^2}} \\ \vdots \\ \theta_K \sqrt{\frac{n_J}{2\sigma_K^2}} \end{pmatrix} , \left(\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ & 1 & \cdots & \rho_{2k} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \bigotimes \begin{pmatrix} 1 & \sqrt{\frac{n_1}{n_2}} & \cdots & \sqrt{\frac{n_1}{n_J}} \\ & 1 & \cdots & \sqrt{\frac{n_2}{n_J}} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \right) \right) \tag{4.7}$$

and

$$\begin{pmatrix} Z_{11} \\ Z_{21} \\ Z_{12} \\ Z_{22} \\ Z_{13} \\ Z_{23} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_1 \sqrt{\frac{n_1}{2\sigma_1^2}} \\ \theta_2 \sqrt{\frac{n_1}{2\sigma_2^2}} \\ \theta_1 \sqrt{\frac{n_2}{2\sigma_1^2}} \\ \theta_2 \sqrt{\frac{n_2}{2\sigma_2^2}} \\ \theta_1 \sqrt{\frac{n_3}{2\sigma_1^2}} \\ \theta_2 \sqrt{\frac{n_3}{2\sigma_2^2}} \end{pmatrix} , \begin{pmatrix} 1 & \rho & \sqrt{\frac{n_1}{n_2}} & \rho\sqrt{\frac{n_1}{n_2}} & \sqrt{\frac{n_1}{n_3}} & \rho\sqrt{\frac{n_1}{n_3}} \\ \rho & 1 & \rho\sqrt{\frac{n_1}{n_2}} & \sqrt{\frac{n_1}{n_2}} & \rho\sqrt{\frac{n_1}{n_3}} & \sqrt{\frac{n_1}{n_3}} \\ \sqrt{\frac{n_1}{n_2}} & \rho\sqrt{\frac{n_1}{n_2}} & 1 & \rho & \sqrt{\frac{n_2}{n_3}} & \rho\sqrt{\frac{n_2}{n_3}} \\ \rho\sqrt{\frac{n_1}{n_2}} & \sqrt{\frac{n_1}{n_2}} & \rho & 1 & \rho\sqrt{\frac{n_2}{n_3}} & \sqrt{\frac{n_2}{n_3}} \\ \sqrt{\frac{n_1}{n_3}} & \rho\sqrt{\frac{n_1}{n_2}} & \sqrt{\frac{n_2}{n_3}} & \rho\sqrt{\frac{n_2}{n_3}} & 1 & \rho \\ \rho\sqrt{\frac{n_1}{n_3}} & \sqrt{\frac{n_1}{n_3}} & \rho\sqrt{\frac{n_2}{n_3}} & \sqrt{\frac{n_2}{n_3}} & \rho & 1 \end{pmatrix} \right) \tag{4.8}$$

4.2.3 Stopping boundaries

In the introduction we explained that the aim of this thesis was to construct a test in such a way as to maintain the family-wise type I error rate and the power in the context of K hypotheses. We have constructed the distribution of the tests in Eq. (4.5), now we need to show that the test controls the FWER in the strong sense and maintains the power. First, we consider the FWER. It is defined through the critical values $\{c_1, ..c_J\}$ described in Eq. (2.45), calculated when $\{Z_1, ..., Z_J\}$ follow the null distribution of Eq. (2.38), but with this setting we do that with K multiple co-primary endpoints using the stopping rules defined in Subsection 1.4.2, step (vi). So for one endpoint, in place of Eq. (2.45), we now define:

$$Pr(\text{stop and reject } H_0 \text{ at or before stage j} \mid \theta = 0) =$$
$$Pr(Z_1 < c_1, \ldots, Z_{j-1} < c_{j-1}, Z_j \geq c_j) = \pi_j/K \tag{4.9}$$

where π_j/K now describes the error rate spending function. Here, the Bonferonni correction is applied as illustrated in Subsection 1.3.3.1.1.

In the setting of K endpoints, in place of Eq. (4.9) we now have:

$$Pr(\text{stop and reject at least one } H_{0k} \text{ at or before stage 1} \mid \theta_k = 0) =$$
$$Pr(Z_{11} > c_1 \text{ or } Z_{21} > c_1 \text{ or} \ldots \text{or } Z_{K1} > c_1 \mid \theta\text{=0}) \leq \pi_1$$

$$Pr(\text{stop and reject at least one } H_{0k} \text{ at or before stage 2} \mid \theta_k = 0) =$$
$$Pr(Z_{11} > c_1 \text{ or } Z_{21} > c_1 \text{ or} \ldots \text{or } Z_{K1} > c_1 \mid \theta\text{=0}) +$$
$$Pr(Z_{11} < c_1, Z_{21} < c_1, \ldots, Z_{K1} < c_1 \text{ , } Z_{12} > c_2 \text{ or}$$
$$Z_{22} > c_2 or \ldots or Z_{K2} > c_2 | \theta = 0) \leq \pi_2$$

$$
\begin{aligned}
&Pr(\text{stop and reject at least one } H_{0k} \text{ at or before stage j} \mid \theta_k = 0) = \\
&Pr(Z_{11} > c_1 \text{ or } Z_{21} > c_1 \text{ or} \ldots \text{or } Z_{K1} > c_1 \mid \theta\text{=0}) + \\
&Pr(Z_{11} < c_1, Z_{21} < c_1, \ldots, Z_{K1} < c_1, Z_{12} > c_2 \text{ or} \\
&Z_{22} > c_2 or \ldots or Z_{K2} > c_2 | \theta = 0) + ... + Pr(Z_{11} < c_1, Z_{21} < c_1, Z_{12} \\
&< c_2, Z_{22} < c_2, \ldots, Z_{K2} < c_2, ..., Z_{1(j-1)} < c_{j-1}, Z_{2(j-1)} < c_{j-1}, \ldots, \\
&Z_{K(j-1)} < c_{j-1}, Z_{1j} > c_j \text{ or } Z_{2j} > c_j \text{ or} \ldots \text{or } Z_{Kj} > c_j \mid \theta_K = 0) \\
&\leq \pi_j
\end{aligned} \tag{4.10}
$$

π_j represents the error spending function at stage j, j = 1,...,J. So, to control the FWER, one must choose c_j to satisfy Eq. (4.9) (i.e, calculate c_j using (4.9) and workout π_j).

Second, we consider the power. It is described as in Eq. (2.41), estimated when $\{Z_1, ..., Z_j\}$ follow the distribution in Eq. (2.38), but here we use K multiple co-primary endpoints. So, in place of Eq. (2.41) the power is now given by:

$$
\begin{aligned}
&Pr(\text{stop and reject at least one } H_{0k} \text{ at or before stage j} \mid \theta_k = \delta_k) = \\
&P(Z_{11} > c_1 \text{ or } Z_{21} > c_1 \text{ or} \ldots \text{or } Z_{K1} > c_1 \mid \theta_k = \delta_k) + \\
&P(Z_{11} < c_1, Z_{21} < c_1 ,\ldots, Z_{K1} < c_1, Z_{12} > c_2 \text{ or} \\
&Z_{22} > c_2 or \ldots or Z_{K2} > c_2 | \theta_k = \delta_k) + ... + P(Z_{11} < c_1, Z_{21} < \\
&c_1, Z_{12} < c_2, Z_{22} < c_2, \ldots, Z_{K2} > c_2, ..., Z_{1(j-1)} < c_{j-1}, Z_{2(j-1)} < \\
&c_{j-1}, \ldots, Z_{K(j-1)} < c_{j-1}, Z_{1j} > c_j \text{ or } Z_{2j} > c_j \text{ or } Z_{K(j)} > c_j \mid \theta_k = \delta_k) \\
&= 1 - \beta
\end{aligned} \tag{4.11}
$$

where $j = 1, ..., J$. For given values of J, α, β and $\{c_1, ..., c_J\}$, the maximum sample

size can be found that satisfies Eq. (4.11) when $\{Z_{11}, \ldots, Z_{KJ}\}$ follow the distribution in Eq. (4.5). We use mvtnorm package in R to compute multivariate normal probabilities as described in more details in 3.1.4.

4.3 Implementation of the method

This section illustrates how the problem defined in Subsection 4.2.1 can be implemented in practice. The general idea is that when the interim data are collected, the planned sample size is re-calculated based on the estimate of the nuisance parameters and the test statistic, and the timings of the test are adjusted accordingly. Details are given in the following subsections.

4.3.1 Before stage 1

At the design stage, we need to:

S0.1. Fix the maximum number of interim analyses J before the study commences.

S0.2. Determine spacing of analyses and

S0.3. Fix times of the interim analyses i.e., $T_{j_0} = (t_{1_0}, t_{2_0}, ..., t_{j_0})$, j = 1,...,J.

S0.4. Choose the overall significance level α and the target power.

S0.5. Specify the type of spending function to apply and use Eq. (2.46) to calculate π_j (j = 1,...,J), the type I error probabilities for each stage.

S0.6. Guess $\rho_{kk'_0}$ ($k' > k$) and $\sigma^2_{k_0}$.

S0.7. Calculate boundaries (see also Subsection 2.3.3): the program at Appendix C finds the univariate boundaries for each stage. It is the modified version of the program

developed by Proschan et al. (2006). For given values of the previous boundaries $c_{1_0}, ..., c_{(j-1)_0}$ and time points $t_{1_0}, ..., t_{(j-1)_0}$, the program finds the current boundary c_{j_0} that satisfies Eq. (4.9).

S0.8. Use Eq. (4.11) to calculate the maximum sample size $Nmax_0$: the program at Appendix D will compute the initial maximum sample size. For given values of J, α/K, β and c_{j_0}, j = 1,...,J, the maximum sample size can be found by computing the multivariate normal probabilities of Eq. (3.9), that satisfies Eq. (4.11) when Z_{kj} (k=1,...,K and j = 1,...,J) follow Eq. (4.5) as illustrated in more details in Subsection 3.1.4. However, the only difference is that the critical value c is replaced by c_j.

4.3.2 Stage 1

At this stage, interim data for stage 1 I_1, which is a fraction of $Nmax_0$, is used to estimate the correlation $\rho_{kk'_1}$ and the variance $\sigma^2_{k_1}$, which are then used to re-estimate the new maximum sample size $Nmax_1$ or maximum information I_{Nmax_1}. The new sample size $Nmax_1$ is used to calculate the information fraction t_1 at stage 1. t_1 is used to calculate the type I error π_1 allocated to stage 1. π_1 is used to find the boundary c_1 at stage 1. c_1 is then compared to the standardised test statistic Z_{k1}, calculated based on interim data at stage 1, to stop the trial or not. In short, the main point about this step is that based on a fraction of the data at the design stage (before stage 1), we can re-estimate nuisance parameters $\rho_{kk'_0}$ and $\sigma^2_{k_0}$ and modify the same size for stage 1 and, at the same time, we can stop the trial or not. The steps for this stage are illustrated in more detail below:

S1.1. Simulate interim data for stage 1, i.e. $I_1 = t_{1_0} Nmax_0$ observations.

S1.2. Use $t_{1_0} Nmax_0$ observations to estimate $\rho_{kk'_1}$ as in Eq. (1.9).

S1.3. Estimate $\sigma^2_{k_1}$ using the blinded method in Eq. (2.10).

S1.4. Use boundaries calculated at the design stage (before stage 1) to estimate the maximum sample size $Nmax_1$ as in step (S0.8).

S1.5. Calculate information fraction at stage 1 : $t_1 = \frac{I_1}{I_{Nmax1}}$ = $\frac{t_{1_0} Nmax_0}{Nmax_1}$.

S1.6. Use Eq. (2.46) to calculate the type I error π_1 allocated to stage 1, i.e. $\pi_1 = f(t_1)$.

S1.7. Use Eq. (2.50) to find boundary c_{1_1} at stage 1.

S1.8. Use Eq. (4.1) to calculate standardised test statistic Z_{k1}.

S1.9. Accept or reject H_{0k} using the program at appendix E.

S1.10. if $Z_{11} > c_{1_1}$ *or...or* $Z_{k1} > c_{1_1}$, reject H_{0k} and stop the trial.

S1.11. Otherwise, go to stage 2

Before stage 2 begins, some adjustments need to be done to information fractions. At stage 1, we realise that the maximum sample size calculated before stage 1 $Nmax_0$ has changed to $Nmax_1$. This is because we have used the estimated correlation $\widehat{\rho_{kk'_1}}$ and the estimated variance $\widehat{\sigma^2_{k_1}}$ instead of $\rho_{kk'_0}$ and $\sigma^2_{k_0}$. That is why we need to modify the information time to reflect this change, that is T_{j_1} = ($\frac{t_{1_0} Nmax_0}{Nmax_1}$, $\frac{t_{2_0} Nmax_1}{Nmax_1}$, ..., $\frac{t_{J_0} Nmax_1}{Nmax_1}$).

4.3.3 Stage 2

For stage 2, we need to estimate nuisance parameters based on stage 2 data and, at the same time, we need to re-estimate the maximum sample size. The new maximum sample size $Nmax_2$ will be different that $Nmax_1$, because we now are going to use the estimated correlation $\widehat{\rho_{kk'_2}}$ and the estimated variance $\widehat{\sigma^2_{k_2}}$ instead of $\widehat{\rho_{kk'_1}}$ and $\widehat{\sigma^2_{k_1}}$. This change will imply that the boundary at stage 1 c_{1_1} would need to be changed to reflect the change in maximum sample size from $Nmax_1$ to $Nmax_2$. However, we cannot to go back to stage 1

and change c_{1_1} based on the new maximum sample size $Nmax_2$, because we have already used it. We now need to construct stage 2 boundary c_{2_2}, allowing for the fact that we have already used the first boundary c_{1_1} at a different time. That is why it is important to modify the information time to reflect this change before stage 2 begins. The steps for stage 2 are as follow:

S2.1. Simulate interim data for stage 2, i.e. $I_{2_2} = t_{2_0} Nmax_1$ observations.

S2.2. Use $t_{2_0} Nmax_1$ observations to estimate $\rho_{kk'_2}$ as in Eq. (1.9).

S2.3. Estimate $\sigma^2_{k_2}$ using the blinded method in Eq. (2.10) based on $t_{2_0} Nmax_1$ observations.

S2.4. Calculate boundaries based on the time of the interim analysis T_{j_1} as in step (S0.7).

S2.5. Estimate the maximum sample size $Nmax_2$ as in step (S0.8).

S2.6. Calculate information fraction at stage 2: $t_2 = \frac{I_2}{I_{Nmax2}} = \frac{t_{2_0} Nmax_1}{Nmax_2}$.

S2.7. Use Eq. (2.46) to calculate the type I error π_2 allocated to stage 2, i.e. $\pi_2 = f(t_2)$.

S2.8. Use c_{1_1} calculated in step (S1.7) to find boundary c_{2_2} at stage 2 as illustrated in step (S0.7) and Eq. (2.51).

S2.9. Use Eq. (4.1) to calculate standardised test statistic Z_{k2}.

S2.10. Accept or reject H_0 using the program at Appendix E.

S2.11. If $Z_{12} > c_{2_2}$ or...or $Z_{k2} > c_{2_2}$, reject H_{0k} and stop the trial.

S2.12. Otherwise, go to stage J.

4.3.4 Stage J

Using the same rationale as in stage 2, we begin by making some adjustments on information fractions $T_{(j-1)_{j-1}}$ = ($\frac{t_{1_0}Nmax_0}{Nmax_{j-1}}$, $\frac{t_{2_0}Nmax_1}{Nmax_{j-1}}$, ..., $\frac{t_{J_0}Nmax_{j-1}}{Nmax_{j-1}}$). Now steps at stage J are:

SJ.1. Simulate interim data for stage J, i.e. $I_J = t_{J_0}Nmax_{j-1}$ observations.

SJ.2. Use $t_{J_0}Nmax_{j-1}$ observations to estimate $\rho_{kk'_J}$ as in Eq. (1.9).

SJ.3. Estimate σ_J^2 using the blinded method in Eq. (2.10) based on $t_{J_0}Nmax_{j-1}$ observations.

SJ.4. Calculate boundaries based on the time of the interim analysis $T_{(j-1)_{j-1}}$ as in step (S0.7).

SJ.5. Estimate the maximum sample size $Nmax_J$ as in step (S0.8).

SJ.6. Calculate information fraction at stage J: $t_J = \frac{I_J}{I_{NmaxJ}} = \frac{t_{J_0}Nmax_{j-1}}{Nmax_J}$.

SJ.7. Use Eq. (2.46) to calculate the type I error π_J allocated to stage J, i.e. $\pi_J = f(t_J)$.

SJ.8. Use c_{1_1} and c_{2_2} to find boundary c_{J_J} at stage J as illustrated in step (S0.7) and Eq. (2.51).

SJ.9. Use Eq. (4.1) to calculate standardised test statistic Z_{kJ}.

Scenario 1: if $t_J > 1$:

SJ.10. Reject H_0 and stop the trial if $Z_{1J} > c_{J_J}$ or...or $Z_{kJ} > c_{J_J}$ using the program at Appendix E.

SJ.11. Otherwise, stop and accept H_0.

Scenario 2:

SJ.12. If $t_J < 1$, the type I error π_J is less than α, i.e. $\pi_J < \alpha$. This implies that we still have a proportion of α to spend, so we need to go to stage *J* + 1 if H_{0k} is not rejected at stage *J*, that is:

SJ.13. Reject H_0 and stop the trial if $Z_{1k} > c_{J_J}$ or ... or $Z_{kJ} > c_{J_J}$ using the program at Appendix E.

SJ.14. Otherwise, go to stage J + 1.

4.3.5 Stage J + 1

This stage happens in Scenario 2 when H_{0k} has not been rejected and $t_J < 1$. It gives us the possibility to develop two options: either we proceed exactly as above by calculating the information fraction t_{J+1} and the type I error π_{J+1} allocated to stage *J* + 1; or we force the trial to stop by fixing the information fraction $t_{J+1} = 1$ and $\pi_{J+1} = \alpha$. Steps at stage *J* + 1 follow the last option.

As in stage J, we begin by making some adjustments to information fractions $T_{(J)_J}$ = $(\frac{t_{1_0} Nmax_0}{Nmax_J}, \frac{t_{2_0} Nmax_1}{Nmax_J}, ..., \frac{t_{J_0} Nmax_{j-1}}{Nmax_J})$.

S(J+1).1. Simulate data for stage J + 1, i.e. $I_{J+1} = t_{J_0} Nmax_J$ observations.

S(J+1).2. Estimate $\rho_{kk'_{J+1}}$ as in Eq. (1.9) using $t_{J_0} Nmax_J$ observations.

S(J+1).3. Estimate σ^2_{J+1} using the blinded method as in Eq. (2.10) and based on $t_{J_0} Nmax_J$ observations.

S(J+1).4. Calculate boundaries based on $T_{(J)_J}$ as in step (S0.7).

S(J+1).5. Estimate the maximum sample size $Nmax_{J+1}$ as in step (S0.8).

S(J+1).6. The information fraction is set to $t_{J+1} = 1$, which implies that the type I error π_{J+1} allocated to stage J + 1 is equal to α i.e. $\pi_{J+1} = \alpha/K$.

S(J+1).7. Use Eq. (2.52) to find boundary $c_{(J+1)}$ at stage J + 1 as illustrated in step (S0.7).

S(J+1).8. Use Eq. (4.1) to calculate standardised test statistic $Z_{k(J+1)}$.

S(J+1).9. If $Z_{1(J+1)} > c_{(J+1)}$ or $\ldots$ $Z_{k(J+1)} > c_{(J+1)}$, reject H_0 and stop the trial.

S(J+1).10. Otherwise, stop and accept H_0.

4.4 Summary

We have illustrated how to implement a GSD with multiple co-primary endpoints. We have shown that if the sample size calculated before the study begins is below or above the actual estimated sample size, we can modify the actual information fraction and calculate the actual boundary accordingly. If at the next stage we realise that the sample size has changed again (is below or above the actual sample size), we can modify the next information fraction and construct the next boundary allowing for the fact we have already used the boundary at the previous stage. The proposed design allows the user to start with a GSD with *J* stages and end up with a GSD with $> J$ stages.

4.5 Example: Three-stage group sequential designs

Suppose that a clinical trial is to be designed to compare an experimental drug E with a placebo control C. Two co-primary endpoints are considered, i.e. K = 2. Patients are randomised in equal numbers between E and C, and a normal distributed response is observed for each of the endpoints. Suppose the parameters of interest representing the mean differences are $\theta_1 = \theta_2 = 0.5$.

Table 4.1: GSD: Implementation of the method

Stages	Values
Before stage 1	
Significance level α	0.025 (one sided)
Power $1-\beta$	0.8
Endpoints	K = 2
Assume ρ_{12_0}	0.5
Assume $\sigma^2_{1_0}$	1.5
Assume $\sigma^2_{2_0}$	1
Number of stages	3
Assume $t_{1_0}, t_{2_0}, t_{3_0}$	$\frac{1}{3}, \frac{2}{3}, \frac{3}{3}$
Calculate $Nmax_0$	72
Stage 1	
Simulate $t_{1_0} Nmax_0$ data	24
Estimate ρ_{12_1}	0.53
Estimate $\sigma^2_{1_1}$	1.30
Estimate $\sigma^2_{2_1}$	0.94
Estimate $Nmax_1$	240
Info. fraction. : $t_1 = \frac{t_{1_0} Nmax_0}{Nmax_1}$	0.10
Calculate c_1 with t_1	3.78
Calculate Z_{11} and Z_{12}	0.84 and 1.20
Conclude $Z_{11} < c_1$ and $Z_{12} < c_1$	Continue to stage 2
Stage 2	
Simulate $t_{2_0} Nmax_1$ data	160
Estimate ρ_{12_2}	0.49
Estimate $\sigma^2_{1_2}$	1.46
Estimate $\sigma^2_{2_2}$	0.89
Estimate $Nmax_2$	220
Info. fraction : $t_2 = \frac{t_{2_0} Nmax_1}{Nmax_2}$	0.72
Calculate c_2 with t_2	2.84
Calculate Z_{21} and Z_{22}	0.85 and 1.60
Conclude $Z_{21} < c_2$ and $Z_{22} < c_2$	Continue to stage 3
Stage 3	
Simulate $t_{3_0} Nmax_2$ data	220
Estimate ρ_{12_3}	0.45
Estimate $\sigma^2_{1_3}$	1.47
Estimate $\sigma^2_{2_3}$	0.93
Estimate $Nmax_3$	250
Info. fraction : $t_3 = \frac{t_{3_0} Nmax_2}{Nmax_3}$	0.88
Calculate c_3 with t_3	2.40
Calculate Z_{31} and Z_{32}	0.16 and 1.61
Conclude $Z_{31} < c_3$ and $Z_{32} < c_3$	Continue to stage 4
Stage 4	
Simulate $t_{3_0} Nmax_3$ data	250
Estimate ρ_{12_4}	0.48
Estimate $\sigma^2_{1_4}$	1.48
Estimate $\sigma^2_{2_4}$	0.95
Estimate $Nmax_4$	230
Fix the info. fraction : t_4	1
Calculate c_4 with t_4	2.33
Calculate Z_{41} and Z_{42}	0.51 and 2.35
Conclude $Z_{41} < c_4$ and $Z_{42} > c_4$	Stop and accept H_{02}

At the design stage (see Subsection 4.3.1 and Figure 4.1), the values considered are summarized in Table 4.1. We assume (or guess) that the variance for endpoint 1 is $\sigma^2_{1_0} = 1.5$, the variance for endpoint 2 is $\sigma^2_{2_2} = 1$ and the correlation between endpoints is $\rho_{12_0} = 0.5$. A three-stage design (J = 3) is required to test $H_{0k} : \theta_k$ = 0, k = 1,2, with a one-sided test type I error rate of $\alpha = 0.025$ and a power of 1 - β = 0.80 for $\theta = 0.5$. We consider the O'Brien and Fleming's spending function as in Eq. (2.48), the time of interim analyses at t_{j_0} = (1/3, 2/3, 3/3), j = 1,2,3 and apply the Bonferroni correction. The first boundary c_{1_0} is found by using the time t_{1_0} and its associated type I error rate π_1, and also the normal distribution function. For given c_{1_0}, t_{1_0} and the cumulative type I error rate to spend at the current time t_{2_0}, the program in Appendix C finds c_{2_0}. For given c_{1_0}, t_{1_0} ,c_{2_0}, t_{2_0} and the cumulative type I error rate to spend by the current time t_{3_0}, the program in appendix C finds c_{3_0}. The initial maximum sample size $Nmax_0$ = 72 is then calculated as described in more detail in step (S0.8).

At stage 1, the values simulated and estimated are summarized in Table 4.1. We simulate stage 1 data as illustrated in step (S1.1). Based on the interim data at stage 1, $\widehat{\rho}_{12_1}$, $\widehat{\sigma}^2_{k_1}$, k = 1,2 and $Nmax_1$ are estimated following steps (S1.2) - (S1.4) as described in Subsection 4.3.2. Suppose the maximum sample size is found to be $Nmax_1$ = 240 as in Figure 4.1. We then use step (S1.5) to calculate the information fraction $t_1 = 0.10$ and the corresponding type I error π_1 as in step (S1.6) to find the boundary c_{1_1} at stage 1 as described in step (S1.7). The boundary $c_{1_1} = 3.78$ is then compared to the standardised test statistic Z_{k1} to accept or reject H_{0k} as described in steps (S1.8) - (S1.11). Figure 4.1 and Table 4.1 show that H_{0k} is not rejected, therefore we proceed to stage 2.

Before we start stage 2, we need to modify the information time to reflect the change in the maximum sample size from $Nmax_0 = 72$ to $Nmax_1 = 240$ as explained at the end of Subsection 4.3.2. This step gives T_1 in Figure 4.1. The values simulated and estimated at this stage are summarized in Table 4.1. We continue and simulate interim data at stage 2

as in step (S2.1) and Figure 4.1. We then estimate $\widehat{\rho}_{12_2}$ and $\widehat{\sigma}^2_{k_2}$, k = 1,2 as in steps (S2.2) and (S2.3). We use step (S2.4) to calculate the boundaries based on the information time T_1 as in Figure 4.1, we do it by using step (S0.7). The maximum sample size $Nmax_2$ = 220 is estimated as in step (S2.5). We calculate the information fraction $t_2 = 0.72$ at stage 2 as in step (S2.6), the type I error π_2 allocated to stage 2 as in step (S2.7) and the corresponding boundary $c_{2_2} = 2.84$ as in step (S2.8). We use steps (S2.9 - S2.11) to calculate standardised test statistic Z_{k2} and do the test of the null hypotheses H_{0k}. Figure 4.1 and Table 4.1 show that H_{0k} is not rejected, hence we go to stage J = 3, which is supposed to be the final stage.

At stage J = 3, we repeat the same process as in stage 2 to calculate the information time to reflect the change in the maximum sample size from $Nmax_1 = 240$ to $Nmax_2 = 220$. This gives T_2 in Figure 4.1. The values simulated and estimated at this stage are summarized in Table 4.1. We continue and simulate interim data at stage 2 as in step (SJ.1) and Figure 4.1. We then estimate $\widehat{\rho}_{12_3}$ and $\widehat{\sigma}^2_{k_3}$, k = 1,2 using steps (SJ.2) and (SJ.3), and calculate the boundaries and the corresponding maximum sample size $Nmax_3$ = 250 as described in steps (SJ.4), (SJ.5) and Figure 4.1. The information fraction t_3 and the type I error π_3 allocated to stage J = 3 are then calculated using steps (SJ.6) and (SJ.7). We use the same c_{1_1}, c_{2_2} calculated before to find $c_{3_3} = 2.40$ as in step (SJ.8) and use step (SJ.9) to calculate the standardised test statistic Z_{k3}. Figure 4.1 shows that H_{0k} is not rejected. We now need to proceed to the next stage, despite the fact that, at the design stage, the plan was to conduct a three-stage group sequential design. The requirement to go to the next stage is justified by the fact that the information fraction $t_3 = 0.88$ is less than 1. This implies that the type I error π_3 allocated to stage 3 will be less than α/K. This situation is equivalent to scenario 2 in Subsection 4.3.4.

At stage J + 1, i.e. J = 4, the values simulated and estimated are summarized in Table 4.1. We set the information fraction to 1 i.e. $t_4 = 1$, so the type I error π_4 allocated to stage 4 is equal to α, i.e. $\pi_4 = \alpha$. We repeat the same procedure as in stage 3 to calculate the

information time to reflect the change in the maximum sample size from $Nmax_2 = 220$ to $Nmax_3 = 250$. This gives T_3 in Figure 4.1. We then estimate $\widehat{\rho}_{12_4}$ and $\widehat{\sigma}^2_{k_4}$, k = 1,2 using steps (S(J+1).2) and (S(J+1).3), and calculate the boundaries and the corresponding maximum sample size $Nmax_4$ = 230 as described in steps (S(J+1).4), (S(J+1).5), Figure 4.1 and Table 4.1. We use the same c_{1_1}, c_{2_2} and c_{3_3} calculated before to find $c_{4_4} = 2.33$ as in step (S(J+1).7) and use step (S(J+1).8) to calculate the standardised test statistic Z_{k4}. We then proceed as in step (S(J+1).9). If H_{0k} is not rejected, we stop the trial as we do not have any α to spend anymore; otherwise we stop and reject H_{k0} as in Table 4.1.

The logic behind the implementation of the proposed group sequential method is that we start by estimating $\widehat{\sigma_1^2}$ using the usual variance estimate for endpoint 1. We do the same for endpoint 2 and get $\widehat{\sigma_2^2}$ and $\widehat{\rho}$. With all these estimates, we can estimate parameters for stage 2 based on stage 1 data. However, the only thing we do not know is the maximum sample size $Nmax_2$ at stage 2. But now we can choose $Nmax_2$ to give the required power based on stage 1 estimates. That is why we assume constant variance throughout the stages. For instance, if we want to go to stage 3, we now have two stages of data to use to estimate σ^2 and ρ, but actually we ignore the fact that they come from two stages. We then estimate σ^2 and ρ the same way using previous data. We then use Z test to assume that in large sample *t* test can be accommodated to Z test.

Throughout this thesis, we assume that we are estimating the same θ as this may be different for stage 1 and stage 2, but we assume they are the same. We also assume that σ^2 is the same throughout the two stages. So the parameters we are estimating are the same, but the estimate may be different, although we believe they are the same. That is why we can use the *Z* test to compute the maximum sample size, because we believe they are the same across stages.

The program in Appendix E is the group sequential method main program. It stim-

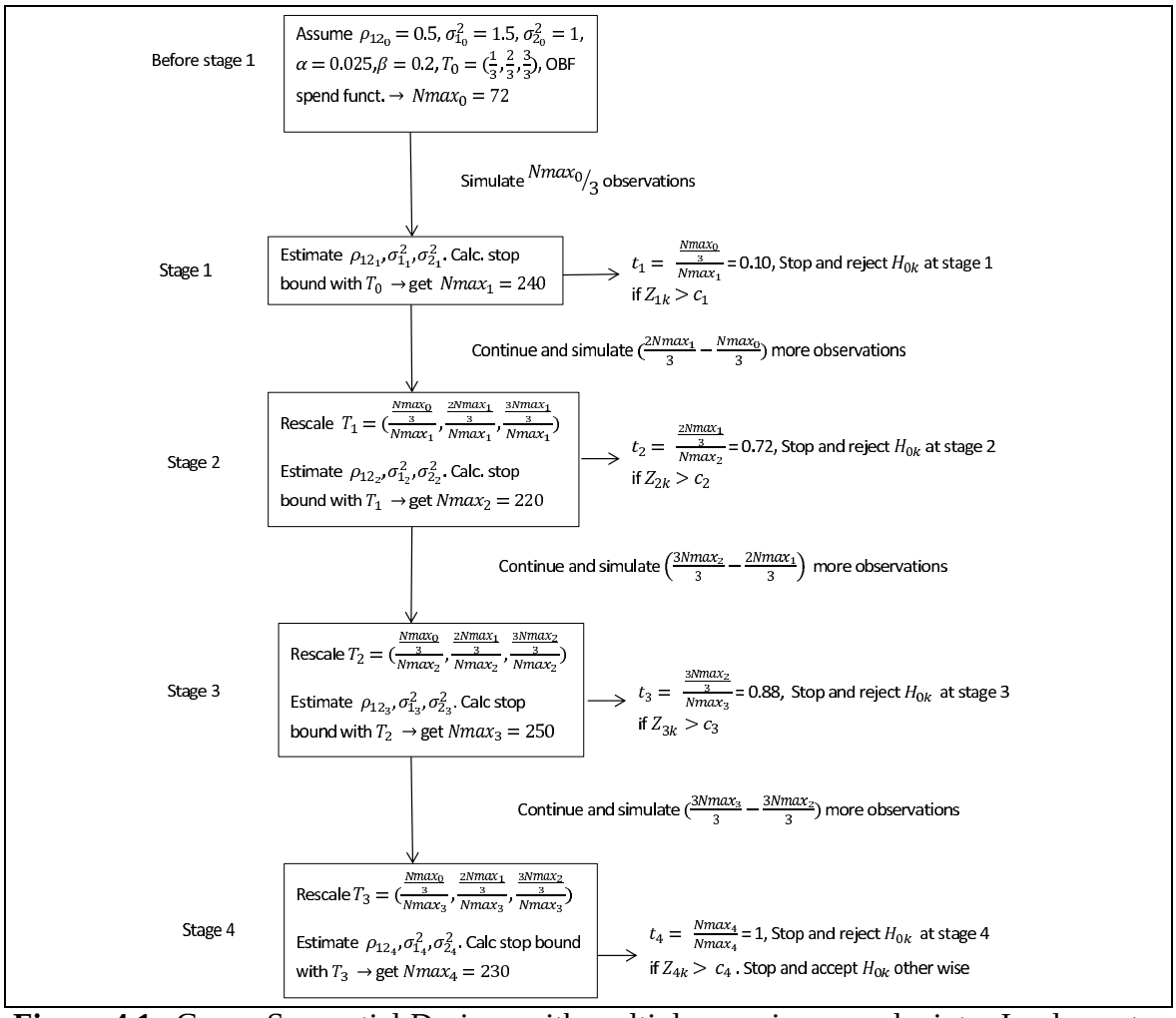

Figure 4.1: Group Sequential Designs with multiple co-primary endpoints: Implementation of the method

ulates the data, calculates the boundary at each stage, estimates the maximum sample size and performs test statistics to either stop the study or not. It uses the program at Appendix D containing the power function defined in Eq. (3.9). It also uses the program at Appendix C which calculates the boundary at each stage as defined in Eq. (2.52). The validation of the main program was performed under different scenarios and the results indicated that the program was working well. For example, we set up some simulations scenarios where we knew in advance the expected results and obtained the anticipated outcomes. One typ-

Table 4.2: Initial values considered in the simulation study.

Fixed parameters	GSD
Significance level α	0.025 (one sided)
Standard error of estimate FWER	0.001
Target power $1-\beta$	0.8
Standard error of estimate power	0.0025
Number of endpoints	K = 2
Number of simulations	100,000
Number of looks = 3	1/3,2/3,3/3
Null hypothesis H_{0k}	$\theta_1 = \theta_2$ = 0
Guessed nuisance parameters	
ρ_{12_0}	0.5
$\sigma^2_{1_0}$	1.5
$\sigma^2_{2_0}$	1

ical trial took 26 to 36 hours to produce the results. The time depends on how the data are correlated and the spending function used.

4.6 Simulation results

As for the two methods developed in Chapter 3, the GSD procedure with multiple endpoints aims to maintain the desired power of the study without inflating the FWER above the nominal level, even if the nuisance parameters $\rho_{kk'}$ and σ^2_k are not known at the planning stage. In this section, we evaluate these characteristics by simulations. We also focused on situations that are typical for phase III trials with two co-primary endpoints. Table 4.2 presents the fixed values considered in the simulation studies. In all the scenarios to be described in Table 4.3 below, we conduct 100,000 simulated trials (standard error of estimate FWER $\alpha = 0.025$ is 0.001 and 0.0025 for the power 1-β = 0.80). We consider the initial guess correlation ρ_{12_0} to be 0.5, the initial guess variance for endpoint 1 $\sigma^2_{1_0}$ to be 1.5 and for endpoint 2 $\sigma^2_{2_0}$ to be 1. We fix in advance the number of looks to 3 and equal spaced i.e, 1/3,2/3,3/3. We aim to randomise patients in equal numbers between E and C.

The scenarios considered in Table 4.3 have the following variable values: In scenario 1, we consider five settings with the same parameters of interest representing the mean difference to be ($\theta_1 = \theta_2 = 0.5$) and we use the O'Brien-Fleming spending function. The five settings have the same values of ρ_{12} ranging from 0 to 1 .i.e. (0, 0.1,. . . , 1) and the same values of true pooled variances for endpoint 1 σ_1^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2). However, each setting has the following true pooled variance for endpoint 2: setting 1, σ_2^2 = 1; setting 2, σ_2^2 = 1.2; setting 3, σ_2^2 = 1.5; setting 4, σ_2^2 = 1.8; and setting 5, σ_2^2 = 2.

In scenario 2, we consider ρ_{12} to be 0.5, pooled variance for endpoint 1 σ_1^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2), pooled variance for endpoint 2 σ_2^2 ranging from 1 to 2 .i.e, (1, 1.2, . . . , 2) and O'Brien-Fleming spending function.

In scenario 3, we consider the same true nuisance parameters as in Setting 3 with different parameters of interest representing the mean difference to be ($\theta_1 = 0.5$, $\theta_2 = 0.7$) and ($\theta_1 = 0.5$, $\theta_2 = 0.7$). We also consider the O'Brien-Fleming spending function.

In scenario 4, we consider the Hwang-Shi-DeCani spending function, with the same true nuisance parameters and the same parameters of interest representing the mean difference as in Setting 3.

4.6.1 FWER, power and sample size in GSD with multiple co-primary endpoints

This subsection aims to check the effect of the mis-specification of the nuisance parameters on the FWER and power using the following settings:

Table 4.3: Scenarios considered in the simulation study.

Variable values	GSD
Scenario 1	**Common values for all settings**
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Type of spending function	O'Brien - Fleming
True nuisance parameters	
ρ_{12}	0,0.1,...,1
σ_1^2	1,1.1,1.2,...,2
Setting 1	
σ_2^2	1
Setting 2	
σ_2^2	1.2
Setting 3	
σ_2^2	1.5
Setting 4	
σ_2^2	1.8
Setting 5	
σ_2^2	2
Scenario 2	Constant ρ_{12}
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Type of spending function	O'Brien - Fleming
True nuisance parameters	
ρ_{12}	0.5
σ_1^2	1,1.1,1.2,...,2
σ_2^2	1,1.1,1.2,...,2
Scenario 3	**Different effect sizes**
Type of spending function	O'Brien - Fleming
Alternative hypothesis	$\delta_1 = 0.5$, $\delta_2 = 0.7$
Alternative hypothesis	$\delta_1 = 0.7$, $\delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	
Scenario 4	**Different type of spending function**
Type of spending function	Hwang-Shih-DeCani
Same alternative hypothesis as in Setting 3	$\delta_1 = \delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	

4.6.1.1 Scenario 1 : Settings 1 - 5

4.6.1.1.1 Scenario 1 : FWER in Settings 1 - 5

Figure 4.2 presents GSD FWER simulation results with the fixed and variable values defined in Table 4.2 and Table 4.3 respectively. It shows that the method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12}, σ_1^2 and σ_2^2. However, we observe the inflation of the FWER in settings 1 2, 3, 4 and 5 for $\rho_{12} = 0$. In setting 1, the FWER has a minimum value of 0.01267 for perfectly correlated data i.e. $\rho_{12} = 1$ and a maximum value of 0.02687 for uncorrelated data, i.e. $\rho_{12} = 0$. In setting 2, a minimum value of 0.01295 for $\rho_{12} = 1$ and a maximum value of 0.02699 for $\rho_{12} = 0$ have been observed. In setting 3, the FWER has a minimum value of 0.01263 for $\rho_{12} = 1$ and a maximum value of 0.02666 for $\rho_{12} = 0$. In setting 4, a minimum value of 0.01301 for $\rho_{12} = 1$ and a maximum value of 0.02573 for $\rho_{12} = 0$ have been observed. Finally, in setting 5, the FWER has a minimum value of 0.01317 for $\rho_{12} = 1$ and a maximum value of 0.02630 for $\rho_{12} = 0$.

4.6.1.1.2 Scenario 1 : Sample size in Settings 1 - 5

Figure 4.3 presents GSD sample size simulation results with the fixed and variable values defined in Table 4.2 and Table 4.3 respectively. The results are presented in all five settings of scenario 1. The figure shows that the sample size increases as ρ_{12}, σ_1^2 and σ_2^2 increase. This shows that the method is working as expected.

4.6.1.1.3 Scenario 1 : Power in Settings 1 - 5

Figure 4.4 presents simulation results for the GSD power with the fixed and variable values defined in Table 4.2 and Table 4.3 respectively. The results are presented in all five

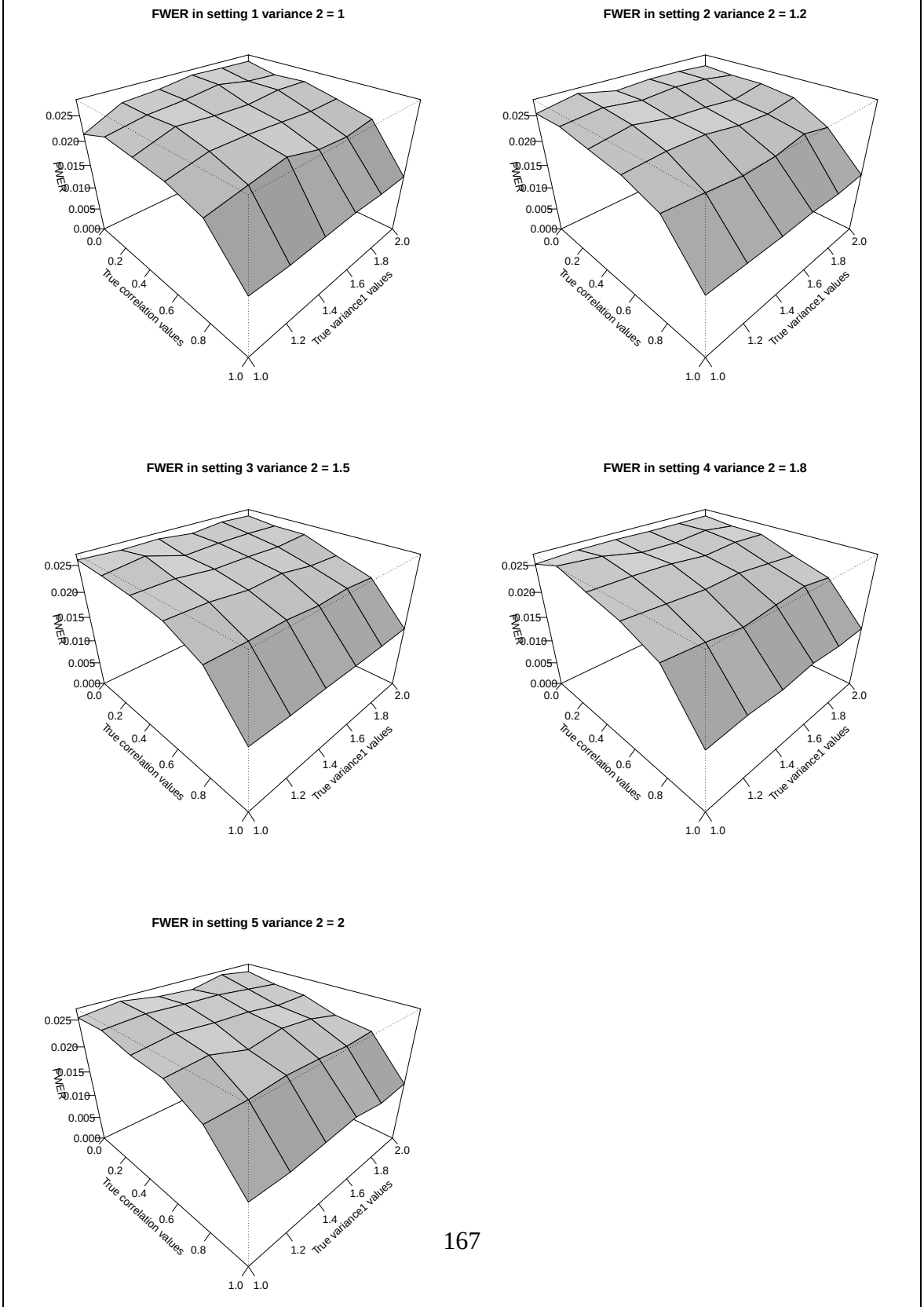

Figure 4.2: GSD FWER in Scenario 1; Settings 1 - 5

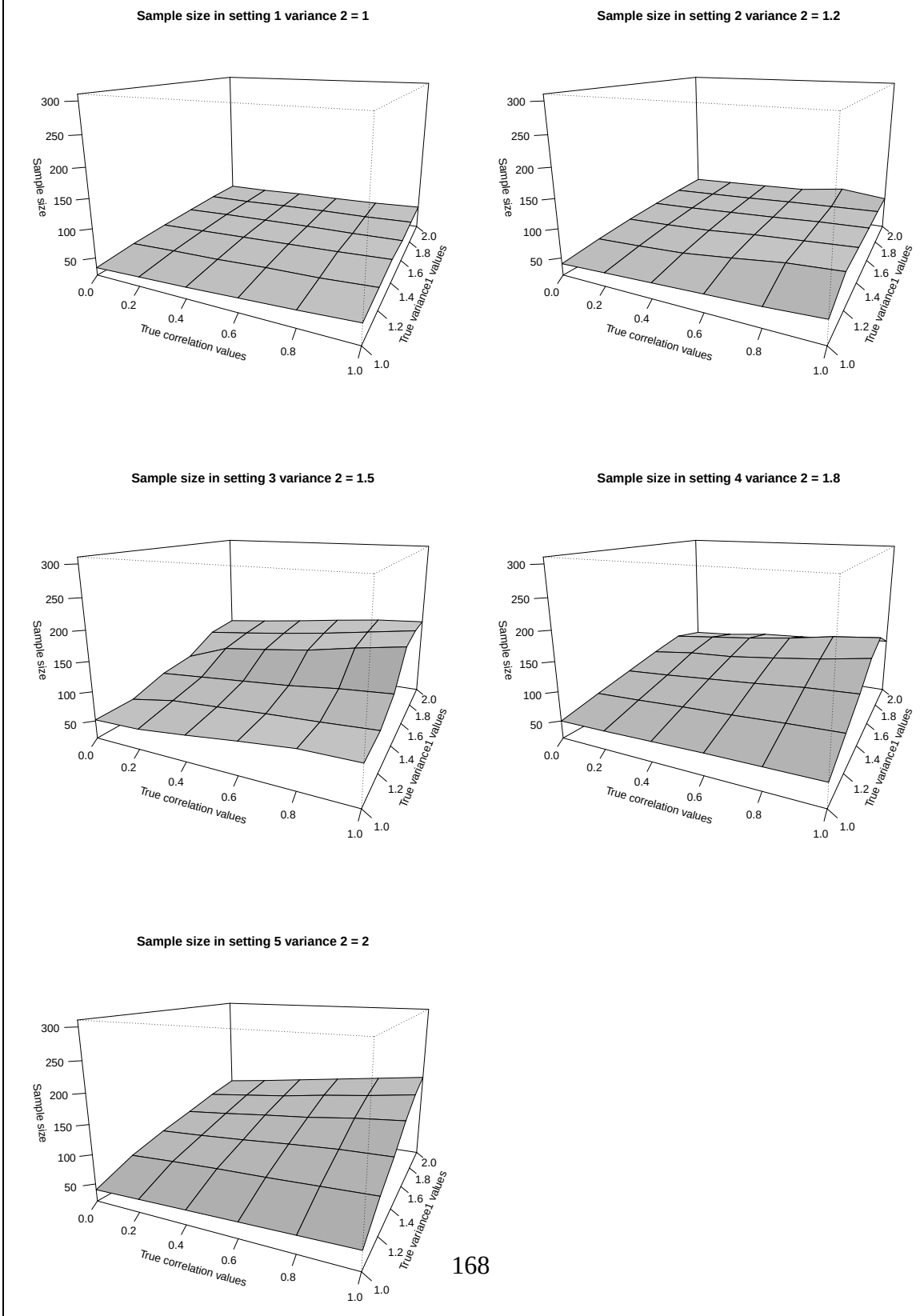

Figure 4.3: GSD Sample size in Scenario 1; Settings 1 - 5

settings of scenario 2. The figure illustrates that the method effectively maintains the power and this is fairly constant despite variation of ρ_{12}, σ_1^2 and σ_1^2. The figure also shows that the power is above the nominal level of 0.80 in some settings.

4.6.1.2 Scenario 2: Constant ρ_{12}

The results in scenario 2 are presented in Figure 4.5. It illustrates that despite variation of σ_1^2 and σ_2^2, the FWER is controlled and fairly constant with the minimum value 0.02223 for σ_1^2 = 1 and σ_2^2 = 1, and the maximum value 0.02504 for σ_1^2 = 1 and σ_2^2 = 2. The same figure illustrates that the sample size increases in the same direction as σ_1^2 and σ_2^2 with the minimum value 48 and the maximum value 158. Finally the figure illustrates that the power is maintained and fairly constant and above the target value despite variation of σ_1^2 and σ_2^2 with 0.7998 the minimum value and 0.8113 the maximum value.

4.6.1.3 Scenarios 1 and 2: Summary and comments on the results

The results in scenario 1 show that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases (Settings 1 - 5). The results in scenario 1 also show that the FWER is above the nominal level of 0.025 for uncorrelated data, i.e. $\rho_{12} = 0$. Nevertheless, this slight increase in the FWER is less than 0.001, hence too small to be practically relevant. Therefore we conclude that for the settings considered here, the GSD procedure controls the FWER. The results in scenario 3 show that the FWER is controlled and fairly constant when ρ_{12} is constant (scenario 3). Again this is know and this shows that the method is working as expected.

The results in scenarios 1 and 2 show that the sample size is increasing in the same direction as ρ_{12}, σ_1^2 or σ_2^2, and the power is fairly constant, and above the target power, despite variations of these nuisance parameters. This is an indication that this method is more powerful than the ones developed in Chapter 3.

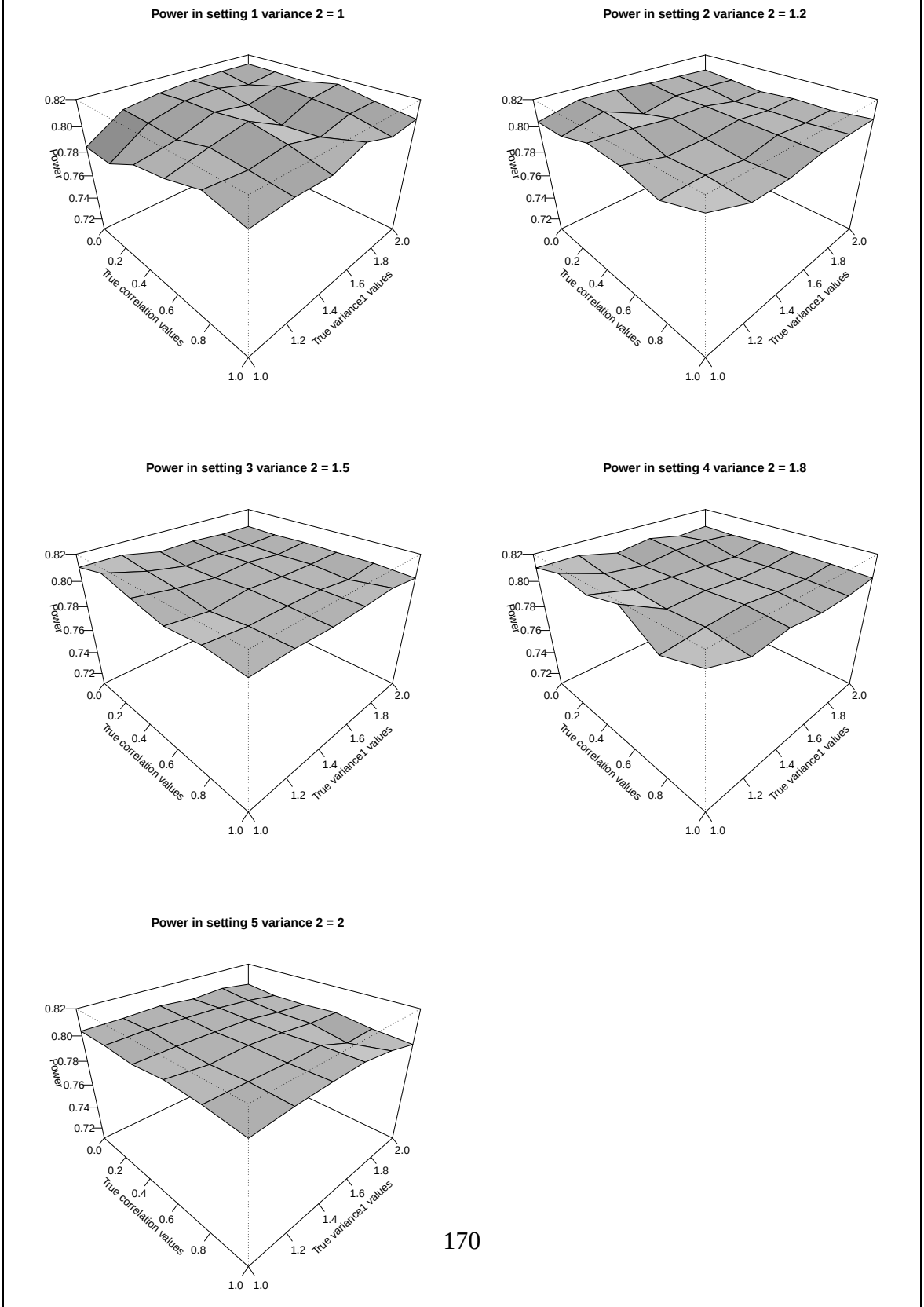

Figure 4.4: GSD Power in Scenario 1; Settings 1 - 5

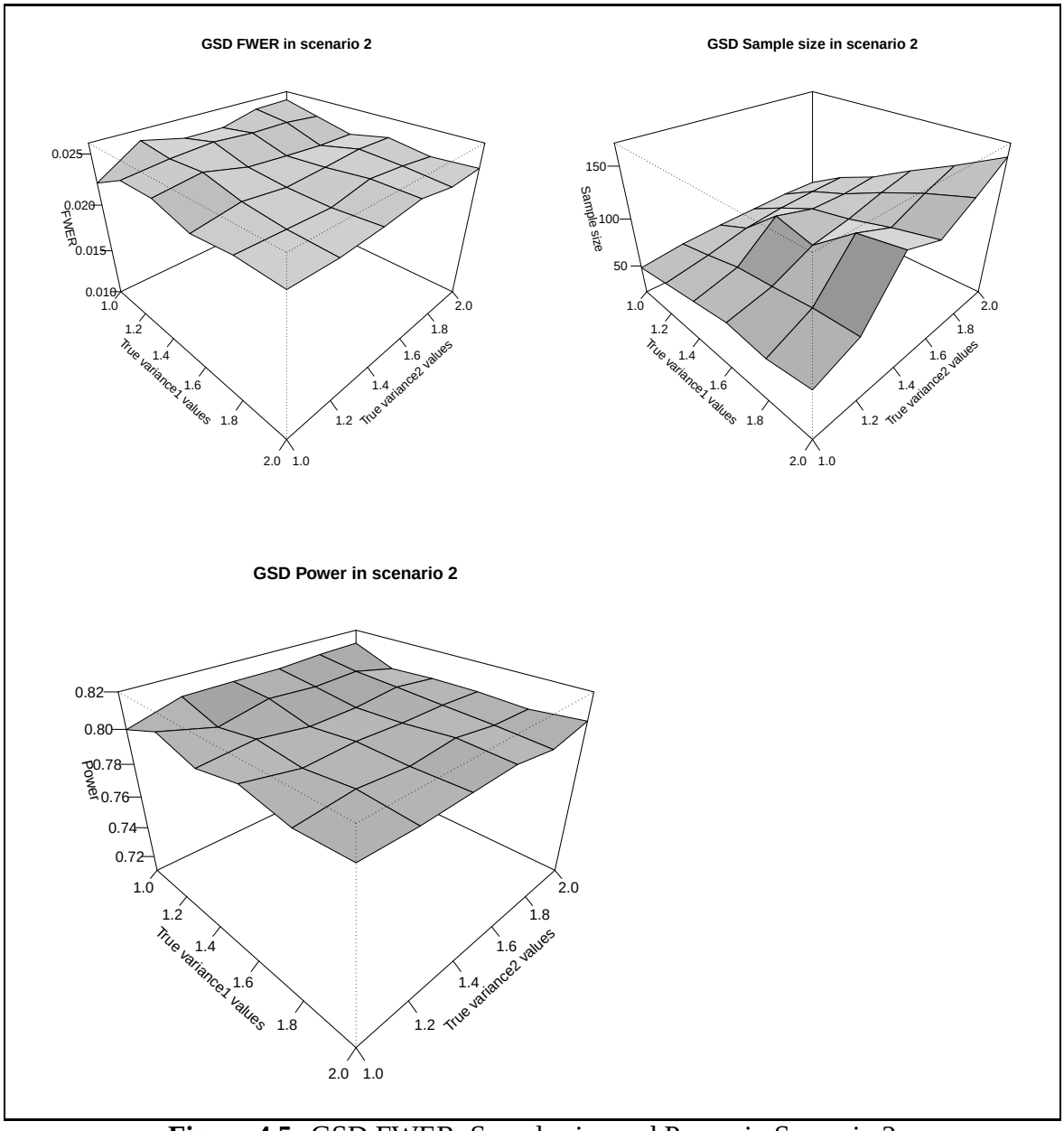

Figure 4.5: GSD FWER, Sample size and Power in Scenario 2

The results in Scenario 1 finally show that all settings control the FWER and maintain the power therefore we only consider setting 3 to check the characteristics of the FWER, power and sample size when different effect sizes and different timings of the interim analysis are considered.

4.6.2 Scenario 3: Different effect sizes

4.6.2.1 Scenario 3: $\delta_1 = 0.5, \delta_2 = 0.7$

Figure 4.6 presents simulation results in scenario 3 with $\delta_1 = 0.5$ and $\delta_2 = 0.7$. It shows that the FWER is controlled at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 with the minimum value 0.01100 for perfectly correlated data and 0.02376 for uncorrelated data. The same figure shows that the sample size decreases as ρ_{12} and σ_1^2 increases with a minimum value of 14 and maximum value of 26. Finally the figure shows that the power is not maintained and decreases when ρ_{12} and σ_1^2 increase with a minimum value of 0.4009 and a maximum value of 0.6756.

4.6.2.2 Scenario 3: $\delta_1 = 0.7, \delta_2 = 0.5$

Figure 4.7 presents a situation where $\delta_1 = 0.7$ and $\delta_2 = 0.5$. It shows that the FWER in this setting is controlled with a minimum value of 0.01100 for $\rho_{12} = 1$ and a maximum value of 0.02429 for $\rho_{12} = 0$. The figure shows that the sample size is decreasing in the opposite direction than ρ_{12} and σ_1^2 with a minimum value of 20 and a maximum value of 46. Finally, the figure shows that the power is not maintained and is decreasing in the opposite direction with ρ_{12} and σ_1^2 with a minimum value of 0.7560 and a maximum value of 0.7879.

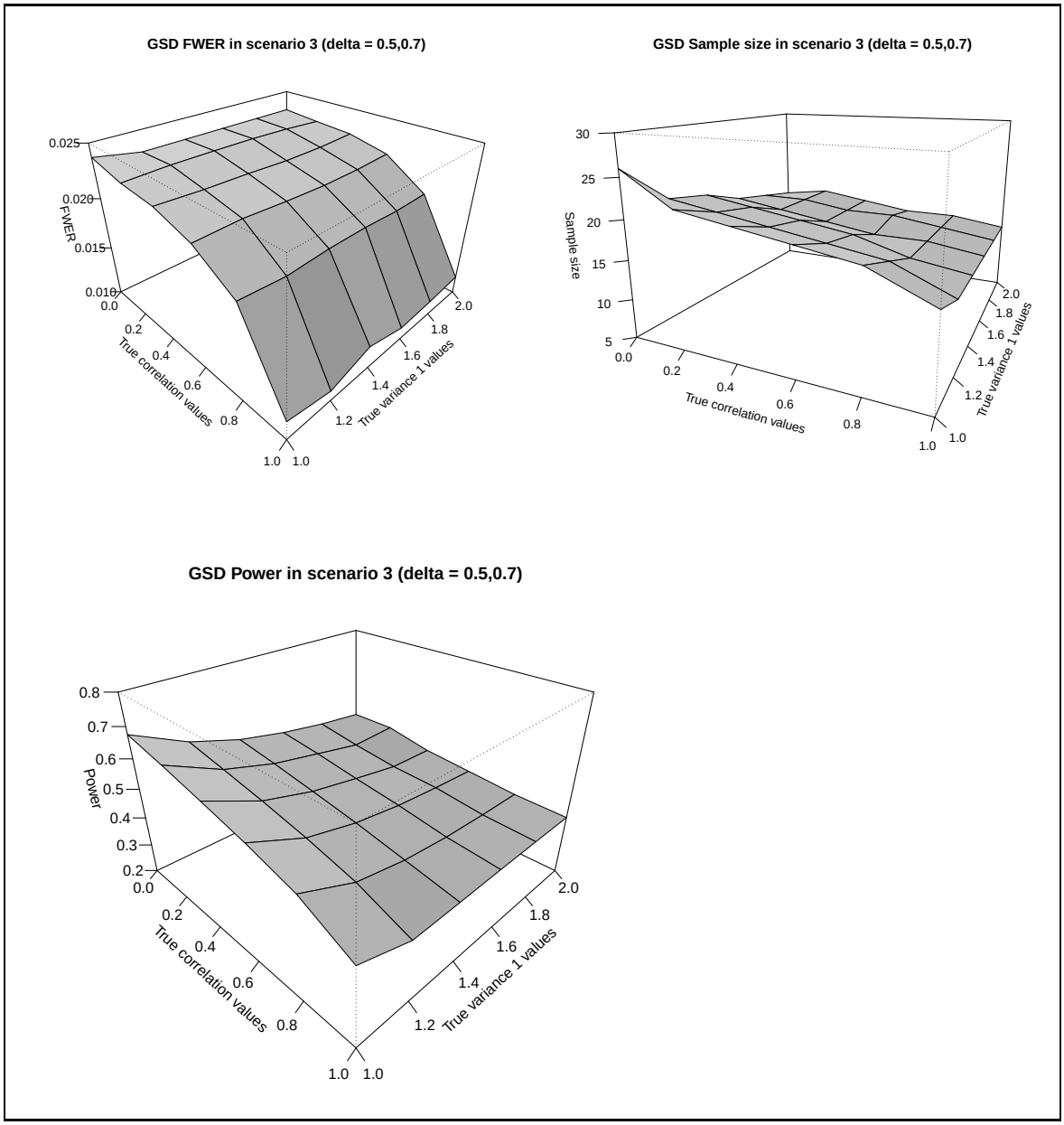

Figure 4.6: GSD FWER, Sample size and Power in Scenario 3 ($\delta_1 = 0.5$, $\delta_2 = 0.7$)

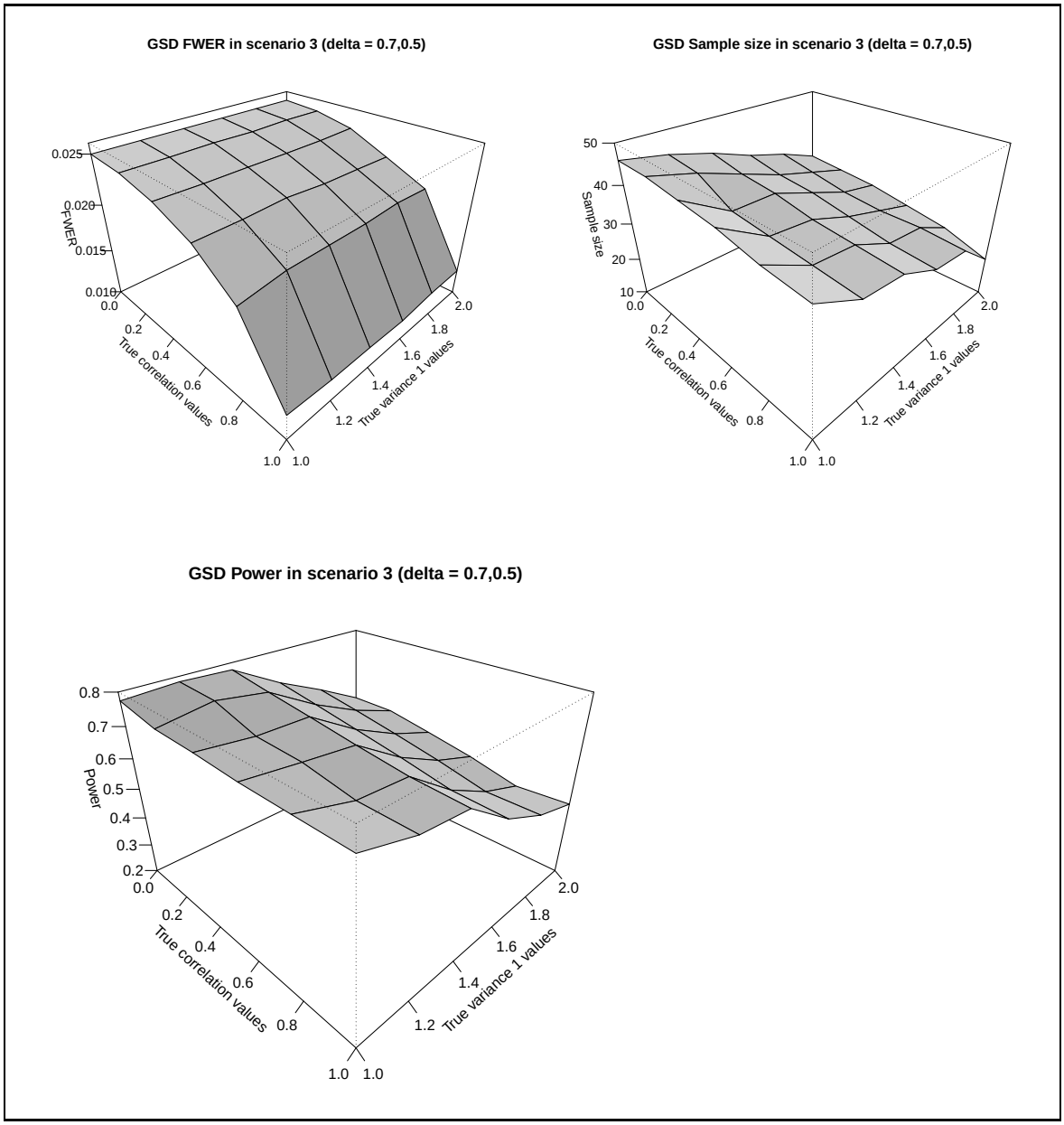

Figure 4.7: FWER, Sample size and Power in Scenario 3 ($\delta_1 = 0.7$, $\delta_2 = 0.5$)

4.6.2.3 Scenario 3: Summary and comments on the results

The results in Scenario 3 show that the FWER is controlled but is conservative as ρ_{12} increases.

The results in Scenario 3 also show that the sample sizes are decreasing in the opposite direction than ρ_{12} and σ_1^2, with a large sample in setting (0.7,0.5) than setting (0.5,0.7). However they (sample sizes) are not large enough to detect different effect sizes in both settings at the same time, hence the reduction in power. This is illustrated in Figures 4.6 and 4.7 as the power decreases in the same direction as the sample size. It is also an illustration that the sample size in this scenario is guided by the big effect size $\delta_1 = 0.7$ than a small one ($\delta_2 = 0.5$), hence it is recommended to use a small effect size for sample size calculation. This is a guarantee that the sample size obtained is large enough to detect even a large effect size and maintain the power.

4.6.3 Scenario 4: Different spending function

4.6.3.1 Scenario 4: Hwang-Shih-DeCani spending function with $\gamma = -10$

Scenario 4 uses the Hwang-Shih-DeCani spending function with $\gamma = -10$. Figure 4.8 presents simulation results with the inputs of Setting 1. It shows that despite variation of ρ_{12} and σ_1^2, the FWER is controlled with a minimum value of 0.01070 and a maximum value of 0.02434. The figure shows that the sample size increases in the same direction as ρ_{12} and σ_2^2 with the minimum value 57 and the maximum value 125. Finally the figure illustrates that the power is not maintained and tends to decease when ρ_{12} and σ_1^2 increase with a minimum value of 0.7627 and a maximum value of 0.7798.

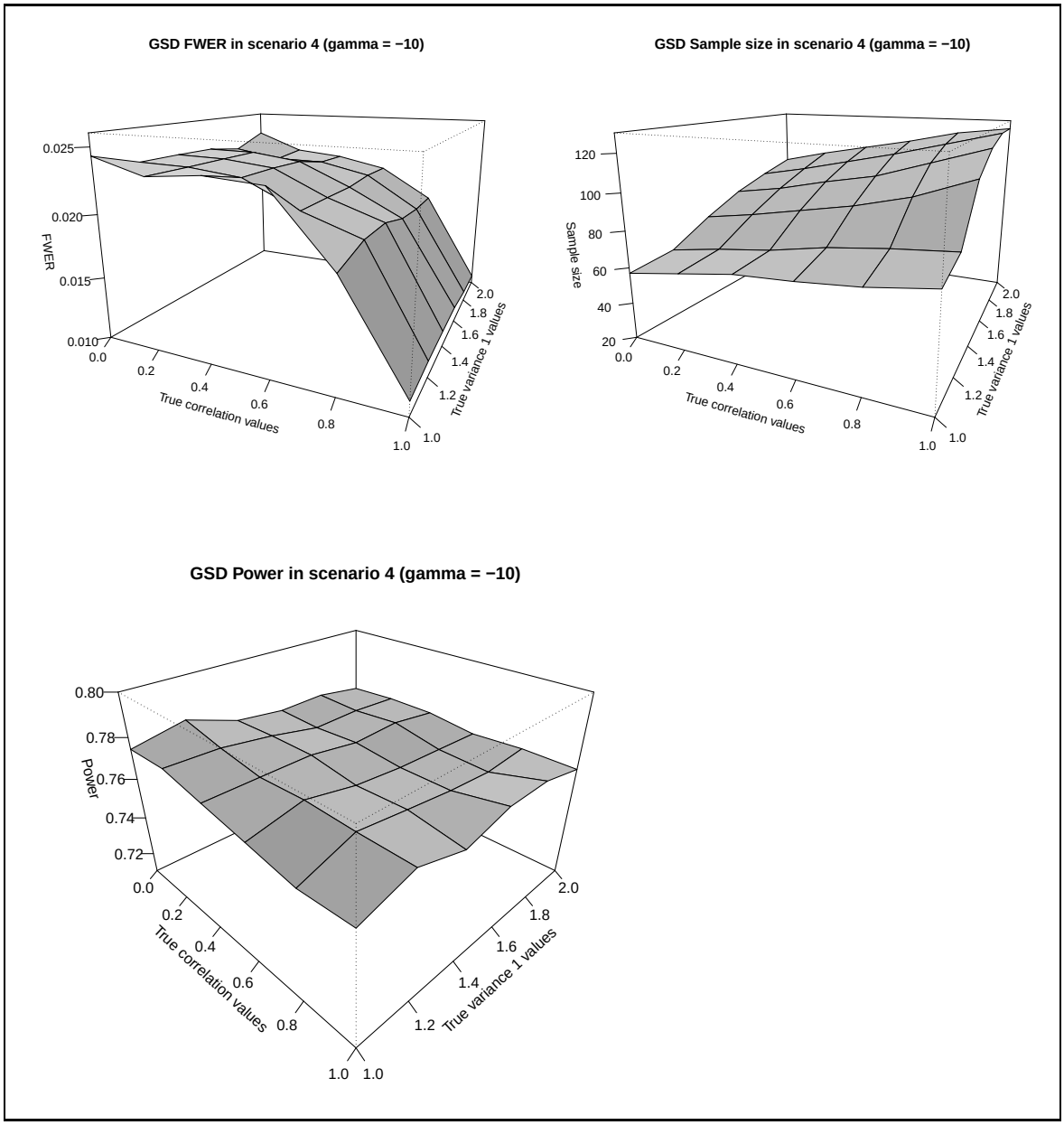

Figure 4.8: GSD FWER, Sample size and Power in Scenario 4 ($\gamma = -10$)

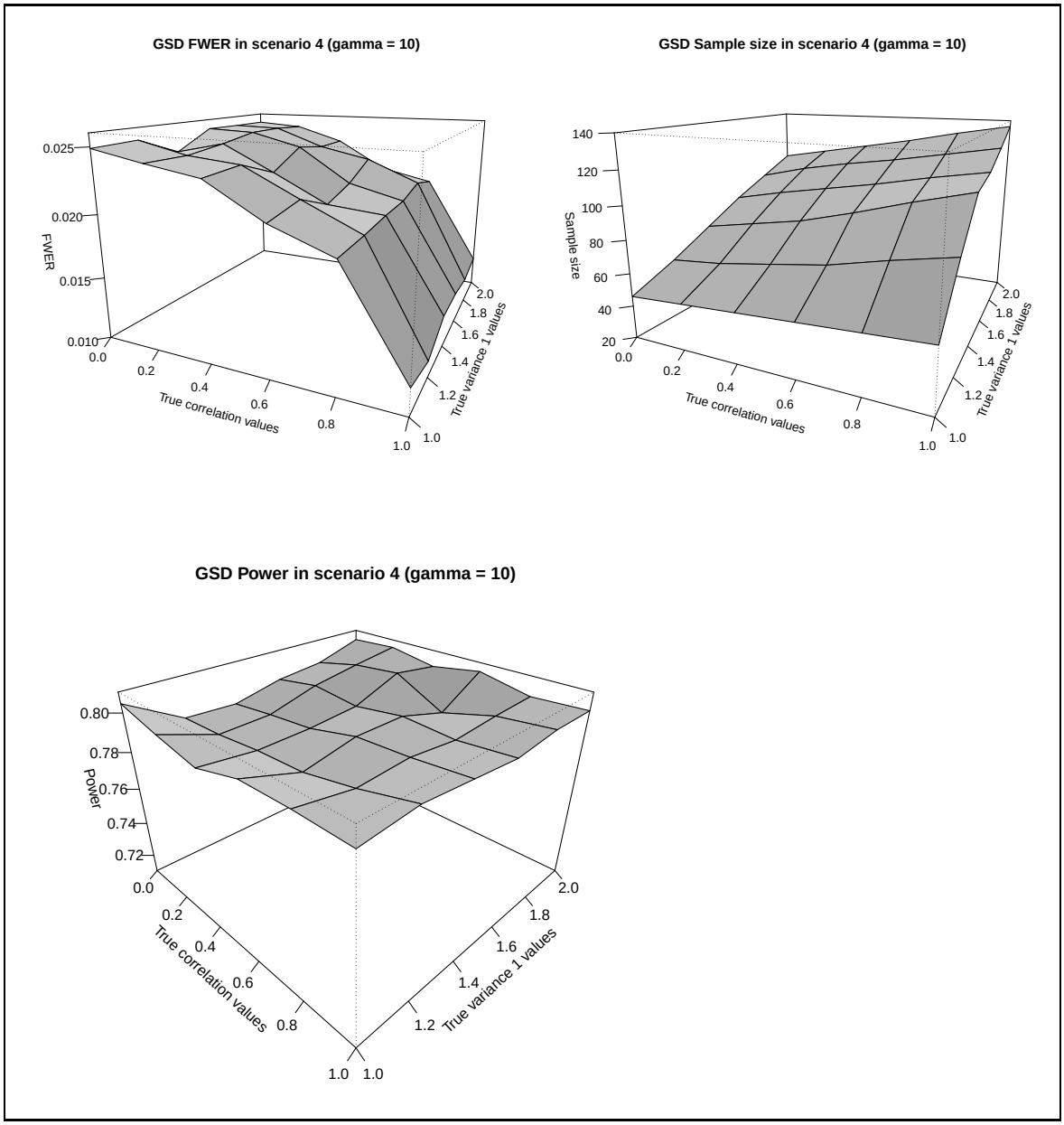

Figure 4.9: GSD FWER, Sample size and Power in Scenario 4 ($\gamma = 10$)

4.6.3.2 Scenario 4: Hwang-Shih-DeCani spending function with $\gamma = 10$

Scenario 4 also uses the Hwang-Shih-DeCani spending function with $\gamma = 10$. Figure 4.9 presents simulation results with the inputs of Setting 1. It shows that despite variation of ρ_{12} and σ_1^2, the FWER is controlled with a minimum value of 0.01170 for $\rho_{12} = 1$ and a maximum value of 0.02515 for $\rho_{12} = 0$. The same figure shows that the sample size increases in the same direction as ρ_{12} and σ_2^2 with the minimum value 46 and the maximum value 136. Finally the figure illustrates that the power is maintained and fairly constant when ρ_{12} and σ_1^2 increase with a minimum value of 0.7877 and a maximum value of 0.8057.

4.6.3.3 Scenario 4: Summary and comments on the results

The results in scenario 4 show that the FWER is controlled but conservative as ρ_{12} increases.

The results in Scenario 4 also show that sample sizes are increasing in the same direction as ρ_{12} and σ_1^2 and are about the same. However, the power for $\gamma = 10$ is maintained compared to the one for $\gamma = -10$. This is because the Hwang-Shih-DeCani spending function with $\gamma = -10$ gives a very conservative spending function. It spends less at the beginning and more later on as illustrated in Figure 2.1. By having this type of spending function, it is likely to go to a late look which means there is no effect of the sample size re-estimation because the trial stops at the first look, consequently the power is reduced despite a big sample size.

4.7 Summary findings from the simulation results

This chapter described group sequential designs in the context of multiple endpoints. The method uses specified stopping rules and a spending function based on the information at each interim analysis. This information is adjusted to allow for the estimated correlation, ρ, between test statistics at each stage. In Section 4.3, we illustrated how to implement this method in practice and gave an example in Section 4.5.

Simulation results presented in Section 4.6 showed that, in most scenarios, the FWER was controlled but became increasingly conservative as ρ_{12} increases. However, we have observed that in Scenario 1 (Settings 1-5), the FWER was above the nominal level of 0.025 for uncorrelated data, i.e. $\rho_{12} = 0$. Nevertheless, this slight increase in the FWER of less than 0.001 was too small to be practically relevant. Therefore we concluded that for the settings considered here, the GSD procedure controls the FWER.

In Section 4.6, we noticed that in most scenarios, the method maintains the power and this was above the target power. However, if strange revision rules are used, the power could not be maintained. For example, we showed in Scenario 3 that the GSD method does not maintain the power when different effect sizes are used simultaneously. Figures 4.6 and 4.7 have shown this. The main findings for this scenario were that the magnitude of the sample size for the GSD method was driven by the big effect size; and this (sample size) was not large enough to detect two different effect sizes at the same time, hence the reduction in power. To reiterate, we recommended to use the small effect size for sample size calculation which is a guarantee that the sample size obtained would be large enough to detect even the large effect size and maintain the power. We also showed in Figure 4.8 that, the power for $\gamma = -10$ was not maintained. This was because the Hwang-Shih-DeCani spending function with $\gamma = -10$ gives a very conservative spending function. By having this type of spending function, it is likely to go to a late look which means there is no effect

of the sample size re-estimation because the trial stops at the first look, consequently the power is reduced despite a big sample size.

In Section 4.6, we finally observed that the sample size was increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2, except in Scenario 3 where it was decreasing as ρ_{12} and σ_1^2 increase. Normally, we would expect the sample size to increase in the same direction as the nuisance parameters in other to maintain the power, but we did not observe this in Scenario 3. That is why the power in this scenario was not maintained.

In the next chapter, we illustrate how the group sequential inverse normal designs, described in Section 2.4, can be extended to the setting of multiple co-primary endpoints.

Chapter 5

Group Sequential Design Inverse Normal Combination tests with multiple co-primary endpoints

This chapter illustrates how the group sequential inverse normal designs, described in Section 2.4, can be extended to the setting of multiple co-primary endpoints. The principle here is the full integration of the concept of inverse normal combination tests illustrated in Section 2.4 into GSD described in Chapter 4. After an introduction in Section 5.1, Section 5.2 presents a framework of analysis. Section 5.3 describes methodology for multiple endpoints. Section 5.4 presents the implementation of the method followed by a worked example in Section 5.5. Section 5.6 presents simulation results.

5.1 Introduction

For the case of early termination for efficacy, we reviewed statistics methods for multiple outcomes in group sequential clinical trials in Section 4.1 and opted to consider multiple hypothesis methods that allow the assessment of differential treatment effects in two or more outcomes. We also reviewed methods allowing for reassessment of the sample size after an interim analysis in group sequential trial with multiple co-primary endpoints in

Chapter 4. The aim of this chapter is to propose a method integrating the concept of inverse normal combination test and multiple co-primary endpoints into group sequential testing. In this setting, the sample size is re-estimated after an interim analysis in a classical group sequential trial, the boundary at each stage is also calculated in the same way, however, the hypothesis testing, assessing the evidence for efficacy of E and C at each stage is conducted using the inverse normal combination test method.

5.2 General framework of analysis

In this section, we define an analysis framework for the GSD inverse normal combination test approach in the context of multiple co-primary endpoints. It is a modified version of the group sequential designs with multiple co-primary endpoints, described in Subsection 1.4.2, but this time, the test statistic is based on the evidence from the different stages of the trial combined via the use of weighted inverse normal functions of the observed *p*-values as described in more detail in Section 2.4. The reason we are doing this is to ensure the FWER control. The general framework is defined in the following setting:

There are two treatments, experimental E and control C.

(i) To reiterate, let X_{kiEj} be the random variable of the k^{th} endpoint for the i^{th} subject in group *E* at stage j ($k = 1, ..., K, i = 1, ..., n_{Ej}, j = 1, ..., J$) and X_{kiCj} be the random variable of the k^{th} endpoint for the i^{th} subject in group *C* at stage j ($k = 1, ..., K, i = 1, ..., n_{Cj}, j = 1, ..., J$).

(ii) We develop one-sided tests as described in Subsection 1.2.5.

(iii) At each stage, we are interested in testing a null hypothesis that two K-dimensional mean vectors of K endpoints are equal against an alternative hypothesis that the difference in mean vectors is a vector of positive K constants:

$$H_{0k} : \theta_k = 0$$

$$H_{1k} : \theta_k > 0.$$

where θ_k is the *k*'th element of θ (a K x 1 column vector of true means) and we are testing a family of *k* hypotheses.

(iv) we use the *p*-value defined in Eq. (3.14) to construct the *p*-value for this setting. Let p_{kj} now denote the *p*-value for endpoint *k* based on *new data* at stage *j* which we write

$$p_{kj} = 1 - \Phi(Z_{kj}) \tag{5.1}$$

where Z_{kj} defined in Eq. (4.1) is now the standardised test statistics for endpoint *k* based on *new data* at stage *j*.

(v) We also use B_j defined in Eq. (3.15) to define the test statistic in the setting of k endpoints and j stages. Suppose B_{kj} now represents the inverse normal combination test statistic for endpoint k at stage *j*.

(vi) Let c_j be the critical value of the accumulating data at stage *j* calculated as in Subsection 2.3.3.

(vii) To reiterate, we assume that X_{ijkE} (X_{ijkC}) has a multivariate normal distribution leading to the multivariate normal distribution for the test statistics B_{kj} with the vector $\theta_k = 0$ and variance $\sigma_k^2 = 1$.

(viii) At each stage *j* we consider the following stopping rules: if $B_{k1} \geq c_1$ or . . . or $B_{k(j-1)} \geq c_{(j-1)}$ or $B_{kj} \geq c_j$, stop at stage *j* and reject H_{0k}, otherwise continue to stage *j* + 1; where c_j represents a critical value at stage *j* defined in step (vi).

(ix) We want to control the FWER in the strong sense, that is, to have
FWER = Pr(reject any true H_{0k}) $\leq \alpha$, under any θ_k, which may combine true and false hypotheses, with at least one true hypothesis.

5.3 Group Sequential Inverse Normal Combination test Designs: Methodology for multiple co-primary endpoints

5.3.1 Definition of the problem

In this subsection, we consider methodology for situations where there are K co-primary correlated endpoints in a clinical trial. The general setting for this problem is defined in Section 5.2. Suppose that E and C are two treatments to be compared in a randomised (phase III) clinical trial with parallel groups. After each group of 2*N* subjects has been randomised in equal numbers to the two therapies and the response obtained, the nuisance parameters are re-estimated based on accumulated data at stage j; the sample size re-estimated and the boundaries adjusted based on re-estimated sample size. However, in contrast to the problem defined in Subsection 4.2.1, the data are now tested using a combination of inverse normal tests of p_{kj}-values defined in Eq. (5.1), that is, at the time of the *j* interim analysis, the decision rule of a group sequential test can be represented as a combination rule for j p_{kj}-values, a series of p_{kj}-values being derived from the data collected before all previous stages and the other p_{kj}-value being derived from the independent new data collected at stage *j*. At each stage of the interim analysis, the further stage of the trial can be understood as an independent new trial. The trial's primary objective is to determine whether E is more efficacious than C in terms of K continuous co-primary responses. This procedure

is conducted at a sequence of up to J interim analyses, each involving a comparison of the evidence for efficacy of E and C, with stopping occurring as soon as one of the interim analyses is in some sense sufficiently convincing.

5.3.2 Test statistics

Suppose that we are interested in repeated looks at the accumulating data on the co-primary endpoints with repeated hypothesis testing. At each interim analysis we will base inference on some calculated test statistics. We need to think about these test statistics and their distributions.

In this subsection, we define test statistics for the setting of K endpoints and J stages, derive distributions and show how this relates to the canonical form as defined in Eq. (2.54) for the setting of a single endpoint.

We use the inverse normal combination test statistic defined in Section 5.2, step (v), to construct our test statistic, and at the design stage, we assume that σ_k^2 is known. Let B_{kj}, k = 1,..,K and j = 1,...,J now denote the inverse normal test for endpoint *k* combining the data as a series of previous p_{kj}-values from before the interim analysis *j* and the new data as the p_{kj}-value at the interim analysis point *j*, which we write as:

$$B_{kj} = w_1^j \Phi^{-1}(1 - p_{k1}) + w_2^j \Phi^{-1}(1 - p_{k2}) + \ldots + w_j^j \Phi^{-1}(1 - p_{kj}) \tag{5.2}$$

where

$$B_{kj} = \sum_{s=1}^{j} w_s^j \Phi^{-1}(1 - p_{ks}) \tag{5.3}$$

where p_{k1} denotes the *p*-value for endpoint *k* based on the data available at stage 1, p_{kj} represents the *p*-value for endpoint *k* based on the new data available at interim analysis *j* and w_s^j denotes the pre-defined weight at each stage satisfying the following equation as suggested by Lehmacher and Wassmer (1999):

$$\sum_{s=1}^{j}(w_s^j)^2 = 1. \tag{5.4}$$

Under H_{0k}, B_{kj} is normally distributed with mean $\theta_k = 0$ and variance $\sum_{s=1}^{j}(w_s^j)^2 = 1$; i.e., $B_{kj} \sim N(0,1)$.

Suppose the covariance between inverse normal combination tests is:

$$Corr(B_{kj}, B_{k'j}) = \rho_{kk'}, k' > k \tag{5.5}$$

so that:

$$Cov(B_{kj}, B_{k'j'}) = w_s^j \rho_{kk'}, k' > k, j' > j, s = 1, \ldots, j. \tag{5.6}$$

As for Eq. (2.54) and under H_{0k}, $B_{11}, \ldots, B_{K1}, \ldots, B_{1J}, \ldots, B_{KJ}$ has a multivariate normal distribution, which we write:

$$\begin{pmatrix} B_{11} \\ \vdots \\ B_{K1} \\ B_{12} \\ \vdots \\ B_{K2} \\ \vdots \\ B_{1J} \\ \vdots \\ B_{KJ} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \begin{pmatrix} (w_1^1)^2 & (w_1^2)^2 & \cdots & (w_1^J)^2 \\ & (w_1^2)^2 + (w_2^2)^2 & \cdots & (w_1^J)^2 + (w_2^J)^2 \\ & & \ddots & \vdots \\ & & & (w_1^J)^2 + \ldots + (w_J^J)^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1K} \\ & 1 & \cdots & \rho_{2K} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \end{pmatrix} \right) \tag{5.7}$$

In the case of K = 2 and J = 3, Eq. (5.7) can be expressed as:

$$\begin{pmatrix} B_{11} \\ B_{21} \\ B_{12} \\ B_{22} \\ B_{13} \\ B_{23} \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} , \left(\begin{pmatrix} (w_1^1)^2 & (w_1^2)^2 & (w_1^3)^2 \\ & (w_1^2)^2 + (w_2^2)^2 & (w_1^3)^2 + (w_2^3)^2 \\ & & (w_1^3)^2 + (w_2^3)^2 + (w_3^3)^2 \end{pmatrix} \bigotimes \begin{pmatrix} 1 & \rho_{12} \\ & 1 \end{pmatrix} \right) \right) \tag{5.8}$$

5.3.3 Stopping boundaries

In the introduction we explained that the aim of this thesis was to construct tests in such a way as to maintain the family-wise type I error rate in the context of K hypotheses.

Suppose that we have the same critical values $\{c_1, ..c_J\}$ calculated as in Subsection 2.3.3. We now need show that using these critical values, the distribution of the inverse normal test statistics constructed under the null hypothesis in Eq. (5.7) controls the FWER in the strong sense. The FWER is defined through the critical values $\{c_1, ..c_J\}$, calculated when $B_{11}, \ldots, B_{K1}, \ldots, B_{1J}, \ldots, B_{KJ}$ follow the null distribution as in Eq. (5.7) and using the stopping rules defined in Section 5.2, step(viii).

So for one endpoint, we define the type I error rate to be

$$\begin{aligned} &Pr(\text{stop and reject } H_0 \text{ at or before stage j} \mid \theta_K = 0) = \\ &Pr(B_1 < c_1, \ldots, B_{j-1} < c_{j-1}, B_j \geq c_j). \end{aligned} \tag{5.9}$$

Now in the setting of K co-primary endpoints, we now have:

Pr(stop and reject at least one H_{0k} at or before stage 1 | $\theta_k = 0$) =

$Pr(B_{11} > c_1$ or $B_{21} > c_1$ | θ=0) $\leq \pi_1$

Pr(stop and reject at least one H_{0k} at or before stage 2 | $\theta_k = 0$) =

$Pr(B_{11} > c_1$ or $B_{21} > c_1$ | θ=0) + $Pr(B_{11} < c_1$, $B_{21} < c_1$, $B_{12} > c_2$ or

$B_{22} > c_2 | \theta = 0) \leq \pi_2$

Pr(stop and reject at least one H_{0k} at or before stage j | $\theta_k = 0$) =

$Pr(B_{11} > c_1$ or $B_{21} > c_1$ | θ=0) + $Pr(B_{11} < c_1$, $B_{21} < c_1$, $B_{12} > c_2$ or

$B_{22} > c_2 | \theta = 0) + ... + Pr(B_{11} < c_1, B_{21} < c_1, B_{12} < c_2, B_{22} < c_2, ...,$

$B_{1(j-1)} < c_{j-1}, B_{2(j-1)} < c_{j-1}, B_{1j} > c_j$ or $B_{2j} > c_j \mid \theta_k = 0)$

$\leq \pi_j$ (5.10)

π_j represents the error spending function at stage j, j = 1,...,J. So, to control the FWER, one must use c_j, calculated as in Subsection 2.3.3, to satisfy Eq. (5.10) (i.e, calculate c_j using (4.9) and workout π_j).

Power consideration is checked by simulations in Subsection 5.6 by using the sample size of the GSD method described in Chapter 4 and the GSD inverse normal combination test statistics described in Subsection 5.3.2. This is because the specification and interpretation of alternative hypotheses is more difficult to define in general as explained by Whitehead (2010).

5.4 Implementation of the method

This section illustrates how the problem defined in Section 5.3.1 can be implemented in practice. The general idea is that when the interim data are collected, the planned sample size is re-calculated based on the estimate of the nuisance parameters and the test statistic and the timing of the test are adjusted accordingly. However, the adjusted test statistic is represented as a combination rule for *j* p_{kj}-values. At each time point of the interim analysis, p_{kj} is independent of p_{k1},...,$p_{k(j-1)}$ so the distribution of the test statistics is unchanged. Furthermore the weights at each look are known and fixed in advance, and they do not depend on the data observed. Details are given in the following subsections.

5.4.1 Design stage

At the design stage, we need to:

W0.1. Fix the maximum number of interim analyses J before the study commences.

W0.2. Determine spacing of analyses and

W0.3. Fix times of the interim analyses i.e., $T_{j_0} = (t_{1_0}, t_{2_0}, ..., t_{j_0})$, j = 1,...,J.

W0.4. Choose the overall significance level α and the target power.

W0.5. Specify the type of spending function to apply and use Eq. (2.46) to calculate π_j (j = 1,...,J), the type I error probabilities for each stage.

W0.6. Guess $\rho_{kk'_0}$ ($k' > k$) and $\sigma^2_{k_0}$.

W0.7. Calculate boundaries as described in more detail in Subsection 4.3.1 step (S0.7).

W0.8. Calculate the maximum sample size $nmax_0$ as specified in more detail in Subsection 3.1.1(or Subsection 4.3.1 step S(0.8)).

W0.9. Fix all weights as they do not depend on the data observed.

5.4.2 Stage 1

At this stage, interim data for stage 1 I_1, which is a fraction of $Nmax_0$, is used to estimate the correlation $\rho_{kk'_1}$ and the variance $\sigma^2_{k_1}$, which are then used to re-estimate the new maximum sample size $Nmax_1$ or maximum information I_{Nmax_1}. The new sample size $Nmax_1$ is used to calculate the information fraction t_1 at stage 1. t_1 is used to calculate the type I error π_1 allocated to stage 1. π_1 is used to find the boundary c_1 at stage 1. c_1 is then compared to the test statistic B_{k1} that results from the inverse normal method of combining independent p_{kj}-values, calculated based on interim data at stage 1, to stop the trial or not. In short, the main point about this step is that based on a fraction of the data at the design stage (before stage 1), we can re-estimate nuisance parameters $\rho_{kk'_0}$ and $\sigma^2_{k_0}$, and modify the sample size for stage 1 and, at the same time, we can stop the trial or not. The steps for this stage are illustrated in more detail below:

W1.1. Simulate interim data for stage 1, i.e. $I_1 = t_{1_0} Nmax_0$ observations.

W1.2. Use $t_{1_0} Nmax_0$ observations to estimate $\rho_{kk'_1}$ as in Eq. (1.9).

W1.3. Estimate $\sigma^2_{k_1}$ using the blinded method in Eq. (2.10), based on $t_{1_0} Nmax_0$ observations.

W1.4. Use boundaries calculated at the design stage (before stage 1) to estimate the maximum sample size $Nmax_1$ as in step (W0.8).

W1.5. Calculate information fraction at stage 1: $t_1 = \frac{I_1}{I_{Nmax1}} = \frac{t_{1_0} Nmax_0}{Nmax_1}$.

W1.6. Use Eq. (2.46) to calculate the type I error π_1 allocated to stage 1, i.e. $\pi_1 = f(t_1)$.

W1.7. Use Eq. (2.50) to find boundary c_{1_1} at stage 1.

W1.8. Use Eq. (1.27) to calculate the degrees of freedom based on the fraction of the data at the design stage (before stage 1) : $df_1 = 2t_{1_0}Nmax_0$-2.

W1.9. Use Eq. (1.23) to calculate *t*-statistics t_{k1} using the variance estimate in step (W1.3).

W1.10. Use Eq. (1.26) to calculate *p*-value p_{k1} from the *t*-distribution.

W1.11. Fix the weight in advance satisfying Eq. (5.4), that is: $(w_1^1)^2 = (\frac{1}{\sqrt{1}})^2 = 1$.

W1.12. Use Eq. (5.3) to calculate the test statistic at stage 1 B_{k1}, calculated based on interim data at stage 1 and the pre-defined weight defined in (W1.11), that is:

$$B_{k1} = \frac{1}{\sqrt{1}}\Phi^{-1}(1 - p_{k1}). \tag{5.11}$$

W1.13. Accept or reject H_{0k} using the stopping rules defined in Section 5.2, step (viii) and implemented in the program at appendix F: if $B_{11} > c_{1_1}$ *or...or* $B_{k1} > c_{1_1}$, reject H_{0k} and stop the trial.

W1.14. Otherwise, go to stage 2.

Before stage 2 begins, some adjustments need to be done to information fractions. At stage 1, we have realised that the maximum sample size calculated before stage 1 $Nmax_0$ has changed to $Nmax_1$. This is because we have used the estimated correlation $\widehat{\rho_{kk_1'}}$ and the estimated variance $\widehat{\sigma_{k_1}^2}$ instead of $\rho_{kk_0'}$ and $\sigma_{k_0}^2$. That is why we need to modify the information time to reflect this change, that is T_{j_1} = $(\frac{t_{1_0}Nmax_0}{Nmax_1}, \frac{t_{2_0}Nmax_1}{Nmax_1}, ..., \frac{t_{J_0}Nmax_1}{Nmax_1})$.

5.4.3 Stage 2

For stage 2, we need to estimate nuisance parameters based on stage 2 data and, at the same time, we need to re-estimate the maximum sample size. The new maximum sample size $Nmax_2$ will be different to $Nmax_1$, because we now are going to use the estimated

correlation $\widehat{\rho_{kk'_2}}$ and the estimated variance $\widehat{\sigma^2_{k_2}}$ instead of $\widehat{\rho_{kk'_1}}$ and $\widehat{\sigma^2_{k_1}}$. This change will imply that the boundary at stage 1 c_1 would need to be changed to reflect the change in maximum sample size from $Nmax_1$ to $Nmax_2$. However, we cannot to go back to stage 1 and change c_1 based on the new maximum sample size $Nmax_2$, because we have already used it. We now need to construct stage 2 boundary c_2, allowing for the fact that we have already used the first boundary c_1 at a different time. That is why it is important to modify the information time to reflect this change before stage 2 begins. The steps for stage 2 are as follows:

W2.1. Simulate interim data for stage 2, i.e. $I_{2_2} = t_{2_0} Nmax_1$ observations.

W2.2. Use $t_{2_0} Nmax_1$ observations to estimate $\rho_{kk'_2}$ as in Eq. (1.9).

W2.3. Estimate $\sigma^2_{k_2}$ using the blinded method as in Eq. (2.10), based on $t_{2_0} Nmax_1$ observations, representing the new data at stage 2.

W2.4. Repeat step (W2.3) but this time using new ($t_{2_0} Nmax_1 - t_{1_0} Nmax_0$) observations for *p*-value calculation.

W2.5. Calculate boundaries based on the time of the interim analysis T_{j_1} as in step (W0.7).

W2.6. Estimate the maximum sample size $Nmax_2$ as in step (W0.8) using correlation estimated in step (W2.2) and variance estimate in step (W2.3).

W2.7. Calculate information fraction at stage 2: $t_2 = \frac{I_2}{I_{Nmax2}} = \frac{t_{2_0} Nmax_1}{Nmax_2}$.

W2.8. Use Eq. (2.46) to calculate the type I error π_2 allocated to stage 2, i.e. $\pi_2 = f(t_2)$.

W2.9. Use c_{1_1} calculated in step (W1.7) to find boundary c_{2_2} at stage 2 as illustrated in step (W0.7) and Eq. (2.51).

W2.10. Use Eq. (1.27) to calculate the degrees of freedom based on the new fraction of data at stage 2: $df_2 = 2(t_{2_0} Nmax_1 - t_{1_0} Nmax_0)$ - 2.

W2.11. Use Eq. (1.23) to calculate *t*-statistics t_{k2} using variance estimate in step (W2.4).

W2.12. Use Eq. (1.26) to calculate *p*-value p_{k2} from the *t*-distribution.

W2.13. Fix the weight in advance satisfying Eq. (5.4), that is: $(w_1^2)^2 + (w_2^2)^2 = (\frac{1}{\sqrt{2}})^2 + (\frac{1}{\sqrt{2}})^2 = 1$.

W2.14. Calculate test statistic based on new data at stage 2, that is:

$$\Phi^{-1}(1 - p_{k2}). \tag{5.12}$$

W2.15. Use Eq. (5.3) to combine test statistics calculated at step (W1.12) and step (W2.14), with pre-defined weight defined in (W2.13):

$$B_{k2} = w_1^2\Phi^{-1}(1 - p_{k1}) + w_2^2\Phi^{-1}(1 - p_{k2}). \tag{5.13}$$

W2.16. Accept or reject H_0 using the program at Appendix F: if $B_{12} > c_{2_2}$ or...or $B_{k2} > c_{2_2}$, reject H_{0k} and stop the trial.

W2.17. Otherwise, go to stage J.

5.4.4 Stage J

Using the same rationale as in stage 2, we begin by making some adjustments to information fractions $T_{(j-1)_{j-1}}$ = $(\frac{t_{1_0}Nmax_0}{Nmax_{j-1}}, \frac{t_{2_0}Nmax_1}{Nmax_{j-1}}, ..., \frac{t_{J_0}Nmax_{j-1}}{Nmax_{j-1}})$. Now steps at stage J are:

WJ.1. Simulate interim data for stage J, i.e. $I_J = t_{J_0}Nmax_{j-1}$ observations.

WJ.2. Use $t_{J_0}Nmax_{j-1}$ observations to estimate $\rho_{kk'_J}$ as in Eq. (1.9).

WJ.3. Estimate $\sigma^2_{k_J}$using the blinded method as in Eq. (2.10), based on $t_{J_0}Nmax_{j-1}$ observations.

WJ.4. Repeat step (WJ.3) but this time using ($t_{J_0}Nmax_{j-1} - t_{(J-1)_0}Nmax_{j-2}$) observations, representing new data at stage J.

WJ.5. Calculate boundaries based on the time of the interim analysis $T_{(j-1)_{j-1}}$ as in step (W0.7).

WJ.6. Estimate the maximum sample size $Nmax_J$ as in step (W0.8) using correlation estimated in step (WJ.2) and variance estimate in step (WJ.3).

WJ.7. Calculate information fraction at stage J: $t_J = \frac{I_J}{I_{NmaxJ}} = \frac{t_{J_0}Nmax_{j-1}}{Nmax_J}$.

WJ.8. Use Eq. (2.46) to calculate the type I error π_J allocated to stage J, i.e. $\pi_J = f(t_J)$.

WJ.9. Use c_{1_1} and c_{2_2} to find boundary c_{J_J} at stage J as illustrated in step (W0.7) and Eq. (2.51).

WJ.10. Use Eq. (1.27) to calculate the degrees of freedom based on the fraction of the data at stage J: $df_J = (t_{J_0}Nmax_{j-1} - t_{(J-1)_0}Nmax_{j-2}) - 2$.

WJ.11. Use Eq. (1.23) to calculate *t*-statistics t_{kJ} using variance estimate in step (WJ.4).

WJ.12. Use Eq. (1.26) to calculate *p*-value p_{kJ} from the *t*-distribution.

WJ.13. Fix the weight in advance satisfying Eq. (5.4), that is::

$$\begin{aligned} & (w_1^J)^2 + (w_2^J)^2 + \ldots + (w_J^J)^2 \\ = \; & (\frac{1}{\sqrt{J}})^2 + (\frac{1}{\sqrt{J}})^2 + \ldots + (\frac{1}{\sqrt{J}})^2 = 1. \end{aligned}$$

WJ.14. Calculate test statistic at stage J, that results from the inverse normal method of combining independent p_{kJ}-value, calculated based on the interim new data at stage J , that is:

$$\Phi^{-1}(1 - p_{kJ}). \tag{5.14}$$

WJ.15. Use Eq. (5.2) to combine test statistics calculated, at step (W1.12.), step (W2.14),..., and step (WJ.14), with pre-defined weight as defined in (WJ.13):

$$\begin{aligned} B_{kJ} &= w_1^J(\Phi^{-1}(1-p_{k1}) + w_2^J\Phi^{-1}(1-p_{k2}) + \dots \\ &\quad + w_J^J\Phi^{-1}(1-p_{kJ}). \end{aligned} \tag{5.15}$$

Scenario 1: if $t_J >= 1$:

WJ.16. Use the program in Appendix F to reject H_{0k} and stop the trial if $B_{1J} > c_{J_J}$ or...or $B_{kJ} > c_{J_J}$.

WJ.17. Otherwise, stop and accept H_{0k}.

Scenario 2: if $t_J < 1$:

WJ.18. If $t_J < 1$, the type I error π_J is less than α, i.e. $\pi_J < \alpha$. This implies that we still have a proportion of α to spend, so we need to go to stage *J* + 1 if H_{0k} is not rejected at stage *J*, that is:

WJ.19. Reject H_{0k} and stop the trial if $B_{k1} > c_{J_j}$ or $\dots$ or $B_{kJ} > c_{J_j}$, using the program in Appendix F.

WJ.20. Otherwise, go to stage J + 1.

5.4.5 Stage J + 1

This stage happens in Scenario 2 when H_{0k} has not been rejected and $t_J < 1$. It gives us the possibility to develop two options: either we proceed exactly as in stage J by calculating

the information fraction t_{J+1} and the type I error π_{J+1} allocated to stage *J* + 1; or we force the trial to stop by fixing the information fraction $t_{J+1} = 1$ and $\pi_{J+1} = \alpha$. We have opted for the last option and the steps for this stage are as follows:

As in stage J, we begin by making some adjustments on information fractions $T_{(J)_J}$ = $(\frac{t_{1_0} Nmax_0}{Nmax_J}, \frac{t_{2_0} Nmax_1}{Nmax_J}, ..., \frac{t_{J_0} Nmax_{j-1}}{Nmax_J})$.

W(J+1).1. Simulate data for stage J + 1, i.e. $I_{J+1} = t_{J_0} Nmax_J$ observations.

W(J+1).2. Use $t_{J_0} Nmax_J$ observations to estimate $\rho_{kk'_{J+1}}$ as in Eq. (1.9).

W(J+1).3. Estimate σ^2_{J+1} using the blinded method as in Eq. (2.10), based on $t_{J_0} Nmax_J$ observations.

W(J+1).4. Calculate boundaries based on the time of the interim analysis $T_{(J)_J}$ as in step (W0.7).

W(J+1).5. Estimate the maximum sample size $Nmax_{J+1}$ as in step (W0.8).

W(J+1).6. The information fraction is set to $t_{J+1} = 1$, which implies that the type I error π_{J+1} allocated to stage J + 1 is equal to α i.e. $\pi_{J+1} = \alpha$.

W(J+1).7. Use Eq. (2.52) to find boundary $c_{(J+1)}$ at stage J + 1 as illustrated in step (W0.7).

W(J+1).8. Use $(Nmax_{J+1} - t_{(J-1)_0} Nmax_{J-1})$ observations instead of $(Nmax_{J+1} - Nmax_J)$ to calculate inverse normal test statistic based on interim data at stage J+1. If we consider $(Nmax_{J+1} - Nmax_J)$ observations, we will not incorporate $Nmax_J - Nmax_{J-1}$ observations which also form part of stage (J+1).

W(J+1).9. Repeat step (W(J+1).3) but this time using $(Nmax_{J+1} - t_{(J-1)_0} Nmax_{J-1})$ observations as explained in step (W(J+1).8).

W(J+1).10. Use Eq. (1.27) to calculate the degrees of freedom based on the fraction of data at stage (J+1): $df_{J+1} = (Nmax_{J+1} - t_{(J-1)_0} Nmax_{J-1}) - 2$.

W(J+1).11. Use Eq. (1.23) to calculate *t*-statistics $t_{k(J+1)}$ using the variance estimate in step (W(J+1).9).

W(J+1).12. Use Eq. (1.26) to calculate *p*-value $p_k(J+1)$ from the *t*-distribution.

W(J+1).13. Fix the weight in advance satisfying Eq. (5.4), that is:

$$\begin{aligned} & (w_1^{(J+1)})^2 + (w_2^{(J+1)})^2 + \ldots + (w_{J+1}^{(J+1)})^2 \\ = \; & (\frac{1}{\sqrt{J+1}})^2 + (\frac{1}{\sqrt{J+1}})^2 + \ldots + (\frac{1}{\sqrt{J+1}})^2 = 1. \end{aligned}$$

W(J+1).14. Calculate the test statistic at stage J+1, that results from the inverse normal method of combining independent $p_{k(J+1)}$-value, calculated based on the interim new data at stage J+1 , that is:

$$\Phi^{-1}(1 - p_{k(J+1)}). \tag{5.16}$$

W(J+1).15. Use Eq. (5.3) to combine the test statistics calculated at step (W1.12.), stage 1 and step (W2.14), stage 2,. . . , step (WJ.14), stage J and step (W(J+1).14), stage (J+1):

$$\begin{aligned} B_{k(J+1)} \; = \; & w_1^{J+1}(\Phi^{-1}(1 - p_{k1}) + w_2^{J+1}\Phi^{-1}(1 - p_{k2}) + \ldots + \\ & w_{J-1}^{J+1}\Phi^{-1}(1 - p_{k(J-1)}) + w_J^{J+1}\Phi^{-1}(1 - p_{kJ}) + \\ & w_{J+1}^{J+1}\Phi^{-1}(1 - p_{k(J+1)}). \end{aligned} \tag{5.17}$$

W(J+1).16. If $B_{k(J+1)} > c_{J+1}$ or . . . or $B_{k(J+1)} > c_{J+1}$, reject H_{0k} and stop the trial.

W(J+1).17. Otherwise, stop and accept H_{0k}.

Table 5.1: GSD: Implementation of the method

Stages	Values
Before stage 1	
Significance level α	0.025 (one sided)
Power $1-\beta$	0.8
Endpoints	K = 2
Assume ρ_{12_0}	0.5
Assume $\sigma^2_{1_0}$	1.5
Assume $\sigma^2_{2_0}$	1
Number of stages	3
Assume $t_{1_0}, t_{2_0}, t_{3_0}$	$\frac{1}{3}, \frac{2}{3}, \frac{3}{3}$
Calculate $Nmax_0$	72
Stage 1	
Simulate $t_{1_0} Nmax_0$ data	24
Estimate ρ_{12_1}	0.53
Estimate $\sigma^2_{1_1}$	1.30
Estimate $\sigma^2_{2_1}$	0.94
Estimate $Nmax_1$	240
Info. fraction. : $t_1 = \frac{t_{1_0} Nmax_0}{Nmax_1}$	0.10
Calculate c_1 with t_1	3.78
Calculate B_{11} and B_{12}	0.84 and 1.20
Conclude $B_{11} < c_1$ and $B_{12} < c_1$	Continue to stage 2
Stage 2	
Simulate $t_{2_0} Nmax_1$ data	160
Estimate ρ_{12_2}	0.49
Estimate $\sigma^2_{1_2}$	1.46
Estimate $\sigma^2_{2_2}$	0.89
Estimate $Nmax_2$	220
Info. fraction : $t_2 = \frac{t_{2_0} Nmax_1}{Nmax_2}$	0.72
Calculate c_2 with t_2	2.84
Calculate B_{21} and B_{22}	0.85 and 1.60
Conclude $B_{21} < c_2$ and $B_{22} < c_2$	Continue to stage 3
Stage 3	
Simulate $t_{3_0} Nmax_2$ data	220
Estimate ρ_{12_3}	0.45
Estimate $\sigma^2_{1_3}$	1.47
Estimate $\sigma^2_{2_3}$	0.93
Estimate $Nmax_3$	250
Info. fraction : $t_3 = \frac{t_{3_0} Nmax_2}{Nmax_3}$	0.88
Calculate c_3 with t_3	2.40
Calculate B_{31} and B_{32}	0.16 and 1.61
Conclude $B_{31} < c_3$ and $B_{32} < c_3$	Continue to stage 4
Stage 4	
Simulate $t_{3_0} Nmax_3$ data	250
Estimate ρ_{12_4}	0.48
Estimate $\sigma^2_{1_4}$	1.48
Estimate $\sigma^2_{2_4}$	0.95
Estimate $Nmax_4$	230
Fix the info. fraction : t_4	1
Calculate c_4 with t_4	2.33
Calculate B_{41} and B_{42}	0.51 and 2.35
Conclude $B_{41} < c_4$ and $B_{42} > c_4$	Stop and accept H_{02}

5.5 Example: Three-stage GSD inverse normal combination test procedure for multiple co-primary endpoints

This section considers the same example as in Section 4.5 to illustrate the GSD inverse normal combination test procedures for multiple co-primary endpoints. Suppose that a clinical trial is to be designed to compare an experimental drug E with a placebo control C. Two co-primary endpoints are considered, i.e. K = 2. Patients are randomised in equal numbers between E and C, and a normally distributed response is observed for each of the endpoints. Suppose the parameters of interest representing the mean differences are $\theta_1 = \theta_2 = 0.5$.

At the design stage (see Subsection 5.4.1 and Figure 5.1), the values considered are summarized in Table 5.1. We assume (or guess) that the variance for endpoint 1 is $\sigma^2_{1_0} = 1.5$, the variance for endpoint 2 is $\sigma^2_{2_2} = 1$ and the correlation between endpoints is $\rho_{12_0} = 0.5$. A three-stage design (J = 3) is required to test $H_{0k} : \theta_k$ = 0, k = 1,2, with a one-sided test type I error rate of $\alpha = 0.025$ and a power of 1 - β = 0.80 for $\theta = 0.5$. We consider the O'Brien and Fleming's spending function as in Eq. (2.48) and the time of interim analyses at t_{j_0} = (1/3, 2/3, 3/3), j = 1,2,3. The first boundary c_{1_0} is found by using the time t_{1_0} and its associated type I error rate π_1, and also the normal distribution function. For given c_{1_0}, t_{1_0} and the type I error rate π_2 to spend at time t_{2_0}, the program in Appendix F finds c_{2_0}. For given c_{1_0}, t_{1_0} ,c_{2_0}, t_{2_0} and the type I error rate to spend by time t_{3_0}, the program in Appendix F finds c_{3_0}. The initial maximum sample size $Nmax_0$ = 72 is then calculated as described in more detail in step (W0.8).

At stage 1, the values simulated and estimated are summarized in Table 5.1. We simulate stage 1 data as illustrated in step (W1.1): $t_{1_0}Nmax_0$ = 27. Based on the interim data at stage 1, $\widehat{\rho}_{12_1}$, $\widehat{\sigma}^2_{k_1}$, k = 1,2 and $Nmax_1$ are estimated following steps (W1.2) -

(W1.4). Suppose the maximum sample size is now $Nmax_1$ = 240 as in Figure 5.1. We then use step (W1.5) to calculate the information fraction t_1 and the corresponding type I error π_1 as in step (W1.6) to find the boundary $c_{1_1} = 3.78$ at stage 1 as described in step (W1.7). Steps (W1.8)-(W1.10) are then used to calculate the degrees of freedom df_1, *t*-test statistics T_{k1} and *p*-value p_{k1} respectively. We fix the weight $w_1^1 = 1$ at stage 1 as in step (W1.11) and calculate the inverse normal test statistics B_{k1} as in step (W1.12). The test statistic B_{k1} is then compared to the boundary c_{1_1} to accept or reject H_{0k} as described in steps (W1.13) - (W1.14). Figure 5.1 and Table 5.1 show that H_{0k} is not rejected, therefore we proceed to stage 2.

Before we start stage 2, we need to modify the information time to reflect the change in the maximum sample size from $Nmax_0 = 72$ to $Nmax_1 = 240$ as explained in Subsection 5.5.3. This step gives T_1 in Figure 5.1. The values simulated and estimated at this stage are summarized in Table 5.1. We continue and simulate interim data at stage 2, $t_{2_0}Nmax_1$ = 160, as in step (W2.1) and Figure 5.1. We then estimate $\widehat{\rho}_{12_2}$ and $\widehat{\sigma}^2_{k_2}$, k = 1,2 as in steps (W2.2) and (W2.3). We use step (W2.5) to calculate the boundaries based on the information time T_1 as in Figure 5.1, we do it by using step (W0.7). The maximum sample size $Nmax_2$ = 220 is estimated as in step (W2.6). We calculate the information fraction $t_2 = 0.72$ at stage 2 as in step (W2.7), the type I error π_2 allocated to stage 2 as in step (W2.8) and the corresponding boundary $c_{2_2} = 2.84$ as in step (W2.9). We use step (W2.4) to estimate the variance based on $(t_{2_0}Nmax_1 - t_{1_0}Nmax_0)$ new observations: $2\frac{Nmax_1}{3} - \frac{Nmax_0}{3}$ = 136. We use new observations at stage 2 to calculate the degrees of freedom df_2, the *t*-test statistics t_{k2} and the *p*-value p_{k2} as in steps (W2.10), (W2.11) and (W2.12) respectively. We then fix the weight $\frac{1}{\sqrt{2}}$ at stage 2 in advance and calculate the test statistics as in steps (W2.13) and (W2.14) respectively. We combine test statistics calculated in step (W1.12) and (W2.13), and calculate the inverse normal test B_{k2} as in step (W2.15), that is: $B_{k2} = \frac{1}{\sqrt{2}}\Phi^{-1}(1 - p_{k1}) + \frac{1}{\sqrt{2}}\Phi^{-1}(1 - p_{k2})$. We then use step (W2.16) to

accept or reject H_{0k}. Figure 5.1 and Table 5.1 show that H_{0k} is not rejected, hence we go to stage J = 3, which is supposed to be the final stage.

At stage J = 3, we repeat the same process as in stage 2 to calculate the information time to reflect the change in the maximum sample size from $Nmax_1 = 240$ to $Nmax_2 = 220$. This gives T_2 in Figure 5.1. The values simulated and estimated at this stage are summarized in Table 5.1. We continue and collect interim data at stage 3, $t_{3_0}Nmax_2$ = 220, as in step (WJ.1). We then estimate $\widehat{\rho}_{12_3}$ and $\widehat{\sigma}^2_{k_3}$, k = 1,2 using steps (WJ.2) and (WJ.3), and calculate the boundaries and the corresponding maximum sample size $Nmax_3$ = 250 as described in steps (WJ.5), (WJ.6) and Figure 5.1 respectively. The information fraction $t_3 = 0.88$ and the type I error π_3 allocated to stage J = 3 are then calculated using steps (WJ.7) and (WJ.8). We use the same c_{1_1}, c_{2_2} calculated before to find $c_{3_3} = 2.40$ as in step (WJ.9). We use step (WJ.4) to estimate the variance based on ($t_{3_0}Nmax_2 - t_{(2)_0}Nmax_1$) new observations: $3\frac{Nmax_2}{3} - 2\frac{Nmax_1}{3}$ = 60. We use new observations at stage 3 to calculate the degrees of freedom df_3, the *t*-test statistics T_{k3} and the *p*-value p_{k3} as in steps (WJ.10), (WJ.11) and (WJ.12) respectively. We then fix the weight $\frac{1}{\sqrt{3}}$ at stage 3 in advance, satisfying $(\frac{1}{\sqrt{3}})^2 + (\frac{1}{\sqrt{3}})^2 + (\frac{1}{\sqrt{3}})^2) = 1$ and calculate the test statistics as in steps (WJ.13) and (WJ.14) respectively. We combine test statistics calculated in step (W1.12), (W2.14) and (WJ.14), and calculate the inverse normal test B_{Jk} as in step (WJ.15), that is: $B_{k3} = \frac{1}{\sqrt{3}}\Phi^{-1}(1 - p_{k1}) + \frac{1}{\sqrt{3}}\Phi^{-1}(1 - p_{k2}) + \frac{1}{\sqrt{3}}\Phi^{-1}(1 - p_{k3})$. Use step (WJ.19) to accept or reject H_{k0}. Figure 5.1 and Table 5.1 show that H_{0k} is not rejected.

We now need to proceed to the next stage, despite the fact that, at the design stage, the plan was to conduct a three-stage GSD inverse normal combination test design. The requirement to go to the next stage is justified by the fact that the information fraction t_3 is less than 1. This implies that the type I error π_3 allocated to stage 3 will be less than α. This situation is equivalent to scenario 2 in Subsection 5.4.4.

At stage (J + 1) = 4, the values simulated and estimated are summarized in Table 5.1. We set the information fraction to 1 i.e. $t_4 = 1$, so the type I error π_4 allocated to stage 4 is equal to α, i.e. $\pi_4 = \alpha$. We repeat the same procedure as in stage 3 to calculate the information time to reflect the change in the maximum sample size from $Nmax_2 = 220$ to $Nmax_3 = 250$. This gives T_3 in Figure 5.1. We then estimate ρ_{12_4} and $\sigma^2_{k_4}$, k = 1,2 using steps (W(J+1).2) and (W(J+1).3) and $t_{3_0} Nmax_3 = 3\frac{250}{3}$ observations, then we calculate the boundaries and the corresponding maximum sample size $Nmax_4$ = 230 as described in steps (W(J+1).4), (W(J+1).5) and Figure 5.1. We use the same c_{1_1}, c_{2_2} and c_{3_3} calculated before to find $c_{4_4} = 2.33$ as in step (W(J+1).7). We also use new observations $(Nmax_4 - t_{(2)_0} Nmax_2)$ = 83 at stage 4 described in step (W(J+1).8) to calculate variance as described in step (W(J+1).9). We then use new observations at stage 4 to calculate the degrees of freedom df_4, the *t*-test statistics t_{k4} and the *p*-value p_{k4} as in steps (W(J+1).10), (W(J+1).11) and (W(J+1).12) respectively. We then fix the weight $\frac{1}{\sqrt{4}}$ at stage 4 in advance, satisfying $(\frac{1}{\sqrt{4}})^2 + (\frac{1}{\sqrt{4}})^2 + (\frac{1}{\sqrt{4}})^2 + (\frac{1}{\sqrt{4}})^2 = 1$ and calculate the test statistics as in steps (W(J+1).13) and (W(J+1).14) respectively. We combine test statistics calculated in step (W1.12), (W2.14), (WJ.14) and (W(J+1).14), and calculate the inverse normal test B_{k4} as in step (W(J+1).15), that is: $B_{k4} = \frac{1}{\sqrt{4}}\Phi^{-1}(1-p_{k1}) + \frac{1}{\sqrt{4}}\Phi^{-1}(1-p_{k2}) + \frac{1}{\sqrt{4}}\Phi^{-1}(1-p_{k3}) + \frac{1}{\sqrt{4}}\Phi^{-1}(1-p_{k4})$. We then proceed as in step (W(J+1).16) to compare B_{k4} and c_{4_4}. If H_{k0} is not rejected, we stop the trial as we do not have any α to spend anymore. Table 5.1 shows that H_{0k} is rejected.

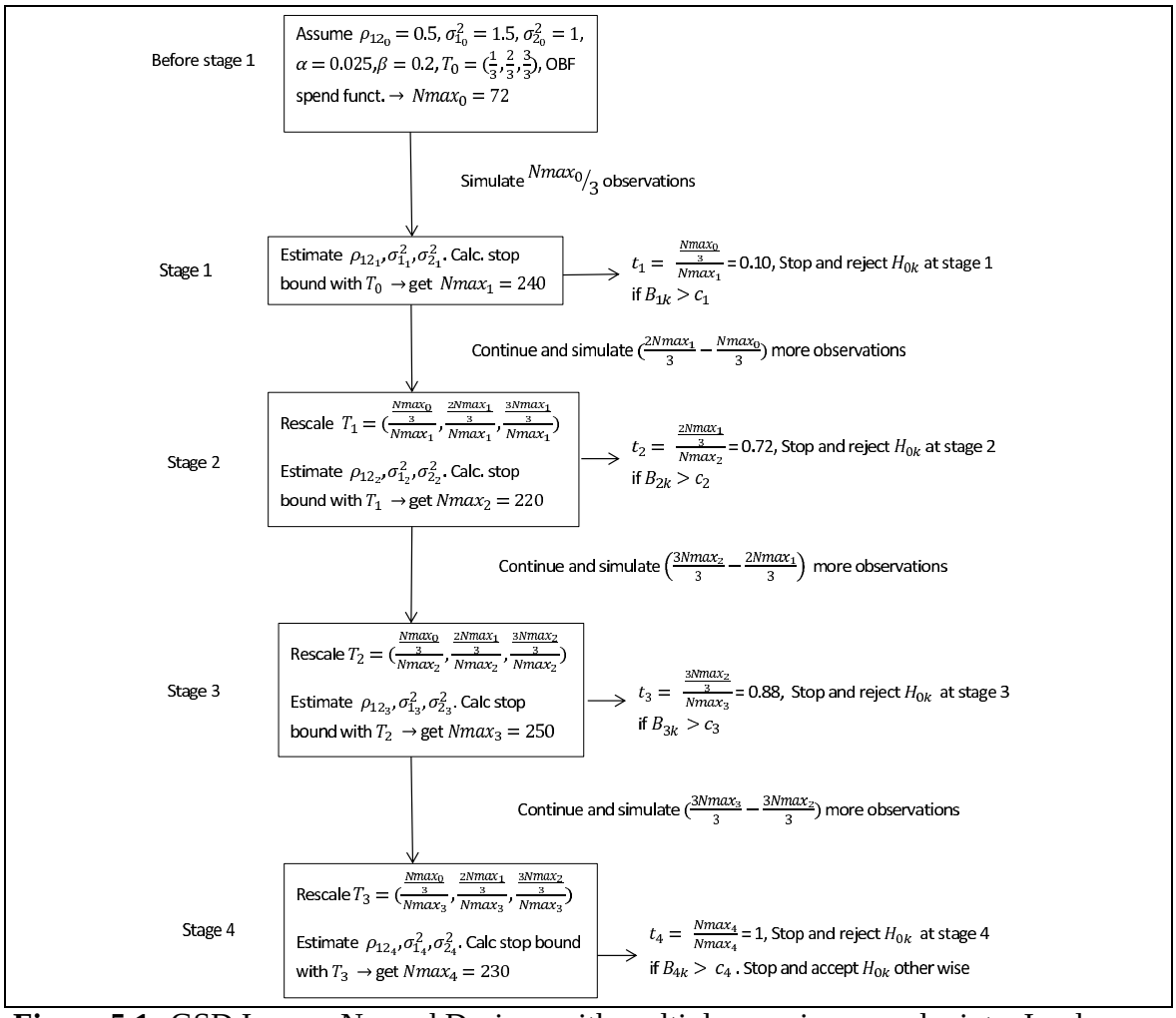

Figure 5.1: GSD Inverse Normal Designs with multiple co-primary endpoints: Implementation of the method

5.6 Simulation results

We explained at the beginning of this chapter that, this method consists of integrating the concept of inverse normal combination tests into GSD with multiple endpoints. We now need to show if using the same settings described in Section 4.6 (and reiterated in Tables 5.2 ad 5.3), we are going to obtain similar results in term of FWER, sample size and power.

Table 5.2: Initial values considered in the simulation study.

Fixed parameters	GSD inverse normal combination test
Significance level α	0.025 (one sided)
Standard error of estimate FWER	0.001
Target power $1-\beta$	0.8
Standard error of estimate power	0.0025
Number of endpoints	K = 2
Number of simulations	100,000
Number of looks = 3	1/3,2/3,3/3
Null hypothesis H_{0k}	$\theta_1 = \theta_2$ = 0
Guessed nuisance parameters	
ρ_{12_0}	0.5
$\sigma^2_{1_0}$	1.5
$\sigma^2_{2_0}$	1

However, here we know that the FWER is controlled, though it may still be conservative. For all the scenarios considered, equal weights for each look are used as illustrated in more detail in the previous section. For each scenario, the OBrien-Fleming spending function is used except for scenario 4 where the Hwang-Shih-DeCani spending function is considered.

5.6.1 FWER, power and sample size in GSD Inverse Normal Combination tests with multiple co-primary endpoints

As for the three methods described previously, this subsection aims to check if the method developed in this chapter controls the FWER and maintains the power if the nuisance parameters change as in the following settings:

5.6.1.1 Scenario 1 : Settings 1 - 5

5.6.1.1.1 Scenario 1 : FWER in Settings 1 - 5

Figure 5.2 presents GSD combination test FWER's simulation results with the fixed

Table 5.3: Scenarios considered in the simulation study.

Variable values	GSD inverse normal combination test
Scenario 1	**Common values for all settings**
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Type of spending function	O'Brien - Fleming
True nuisance parameters	
ρ_{12}	0,0.1,...,1
σ_1^2	1,1.1,1.2,...,2
Setting 1	
σ_2^2	1
Setting 2	
σ_2^2	1.2
Setting 3	
σ_2^2	1.5
Setting 4	
σ_2^2	1.8
Setting 5	
σ_2^2	2
Scenario 2	**Constant** ρ_{12}
Alternative hypothesis $\theta_k = \delta_k$	$\delta_1 = \delta_2 = 0.5$
Type of spending function	O'Brien - Fleming
True nuisance parameters	
ρ_{12}	0.5
σ_1^2	1,1.1,1.2,...,2
σ_2^2	1,1.1,1.2,...,2
Scenario 3	**Different size effects**
Type of spending function	O'Brien - Fleming
Alternative hypothesis	$\delta_1 = 0.5$, $\delta_2 = 0.7$
Alternative hypothesis	$\delta_1 = 0.7$, $\delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	
Scenario 4	**Different type of spending function**
Type of spending function	Hwang-Shih-DeCani
Same alternative hypothesis as in Setting 3	$\delta_1 = \delta_2 = 0.5$
Same true nuisance parameters as in Setting 3	

and variable values defined in Table 5.2 and Table 5.3 respectively. The figure shows that this method effectively controls the overall FWER at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 in all five settings. In setting 1, the FWER has a minimum value of 0.00895 for perfect correlated data i.e. $\rho_{12} = 1$ and a maximum value of 0.02301 for uncorrelated data, i.e. $\rho_{12} = 0$. In setting 2, a minimum value of 0.00904 for $\rho_{12} = 1$ and a maximum value of 0.02129 for $\rho_{12} = 0$ have been observed. In setting 3, the FWER has a minimum value of .00961 for $\rho_{12} = 1$ and a maximum value of 0.02301 for $\rho_{12} = 0$. In setting 4, a minimum value of 0.00943 for $\rho_{12} = 1$ and a maximum value of 0.02325 for $\rho_{12} = 0$ have been observed. Finally, in setting 5, the FWER has a minimum value of 0.00892 for $\rho_{12} = 1$ and a maximum value of 0.02339 for $\rho_{12} = 0$.

5.6.1.1.2 Scenario 1 : Sample size in Setting 1 - 5

Figure 5.3 presents GSD combination test sample size simulation results with the fixed and variable values defined in Table 5.2 and Table 5.3 respectively. The results are presented in all five settings of scenario 1. The figure shows that the sample size increases as ρ_{12}, σ_1^2 and σ_2^2 increase. Again, this is an indication that the method is working.

5.6.1.1.3 Scenario 1 : Power in Settings 1 - 5

Figure 5.4 presents simulation results for the power with the fixed and variable values defined in Table 5.2 and Table 5.3 respectively. The figure illustrates that the method does not maintain the power in all five settings.

5.6.1.2 Scenario 2 : Constant ρ_{12}

The results in scenario 2 are presented in Figure 5.5. Its illustrates that despite variation of σ_1^2 and σ_2^2, the FWER is controlled with small values for σ_1^2 = 1.2 and σ_2^2 = 1, and fairly

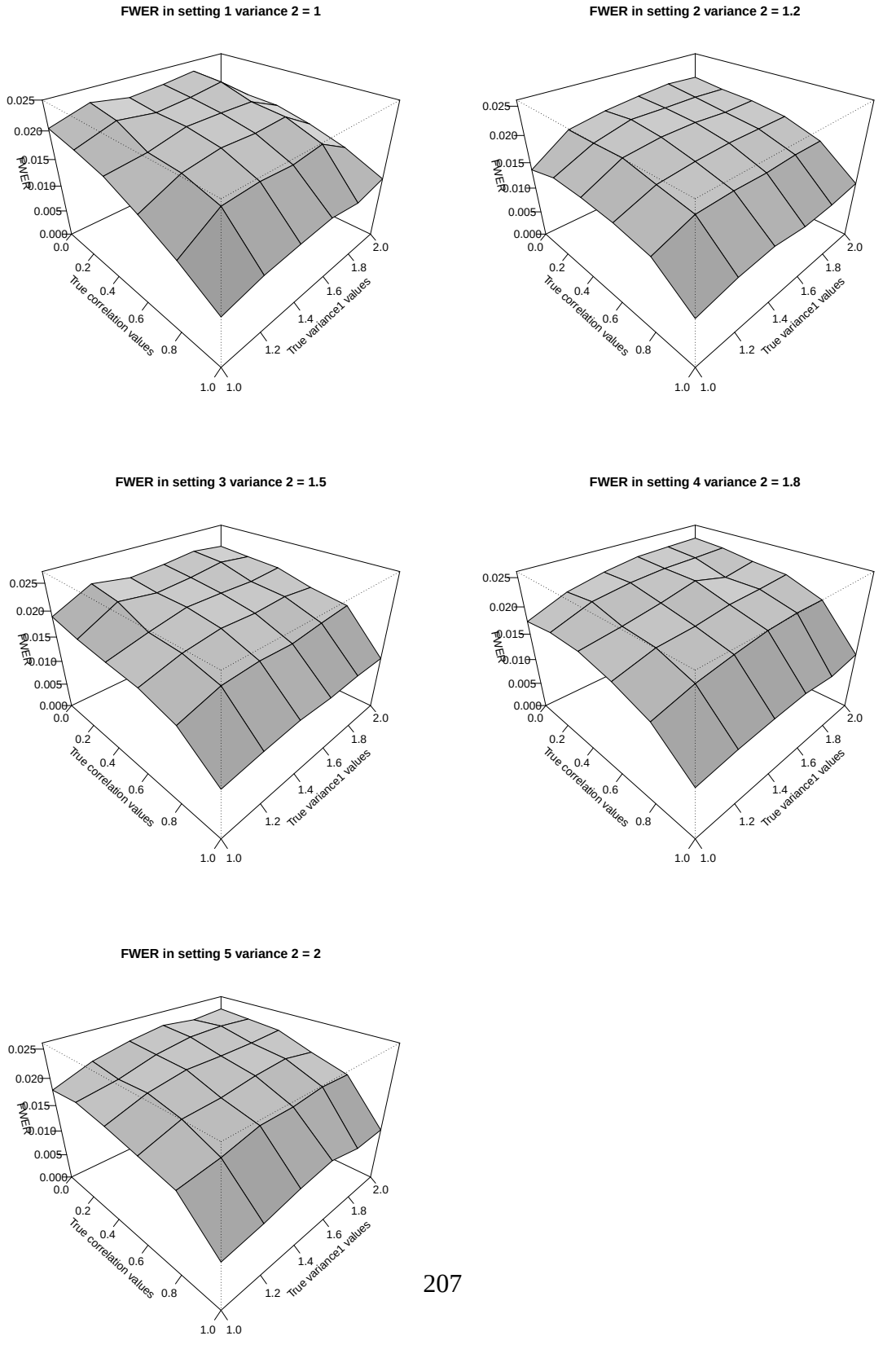

Figure 5.2: GSD combination FWER in Scenario 1; Settings 1 - 5

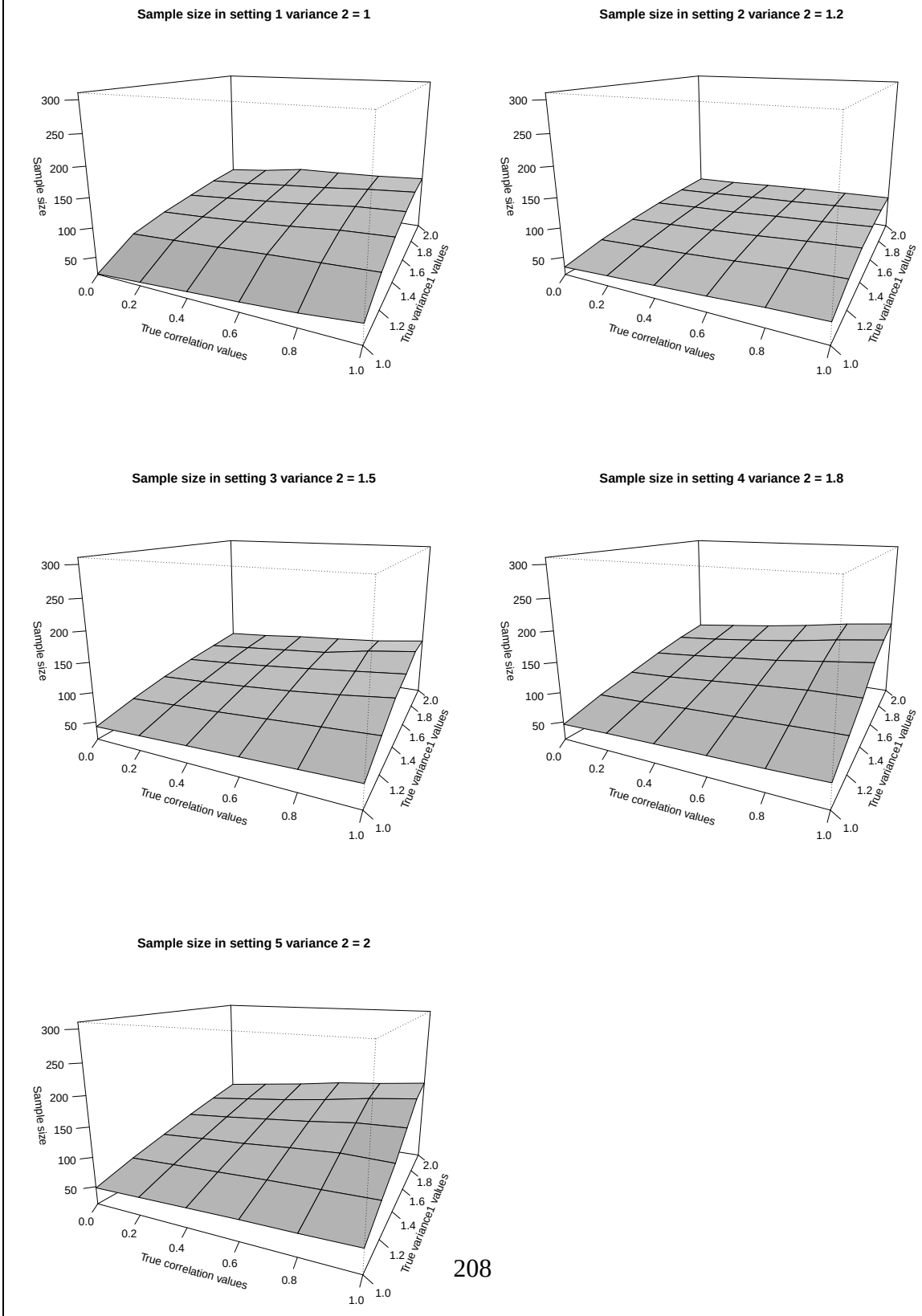

Figure 5.3: GSD combination Sample size in Scenario 1; Settings 1 - 5

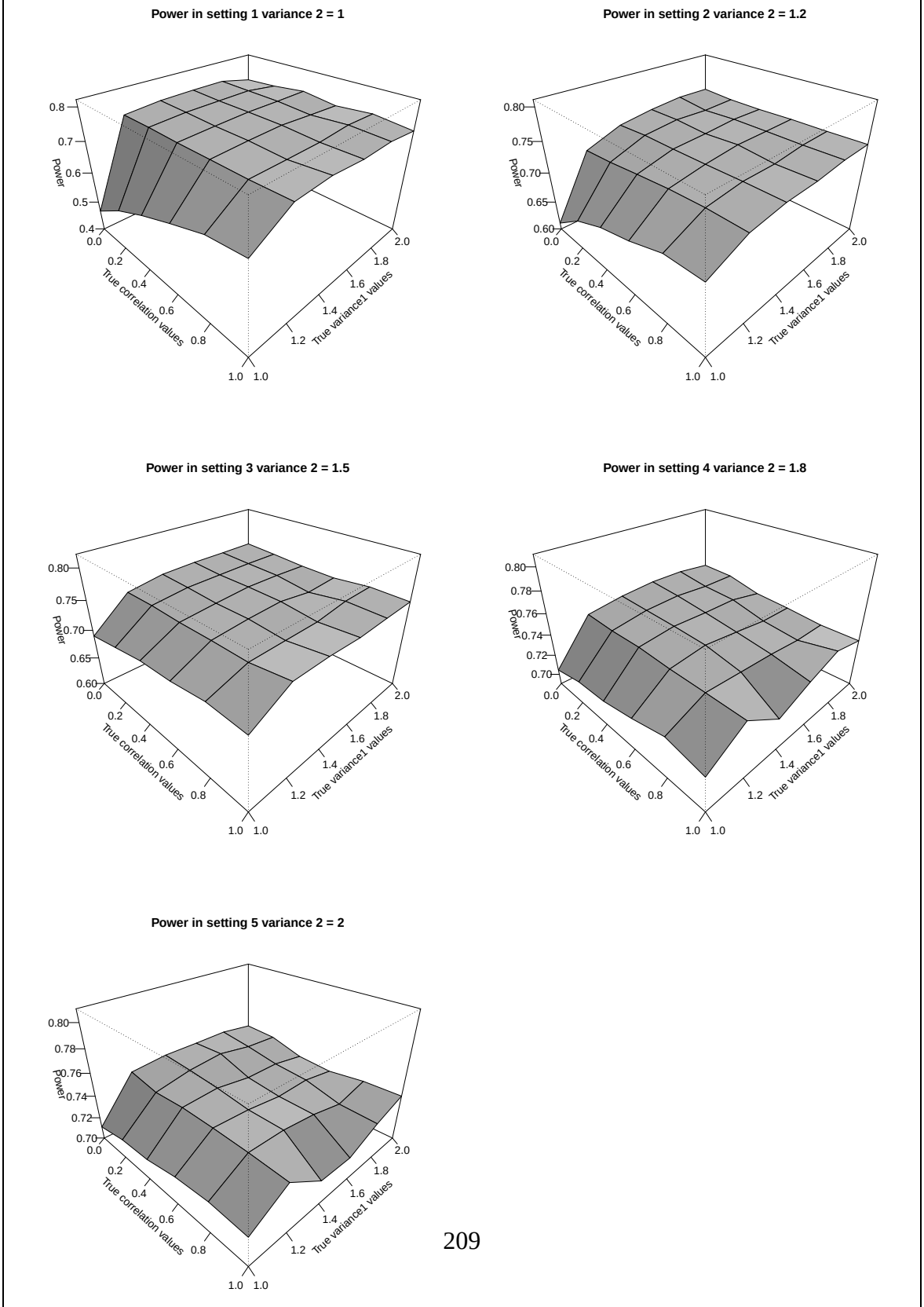

Figure 5.4: GSD combination Power in Scenario 1; Settings 1 - 5

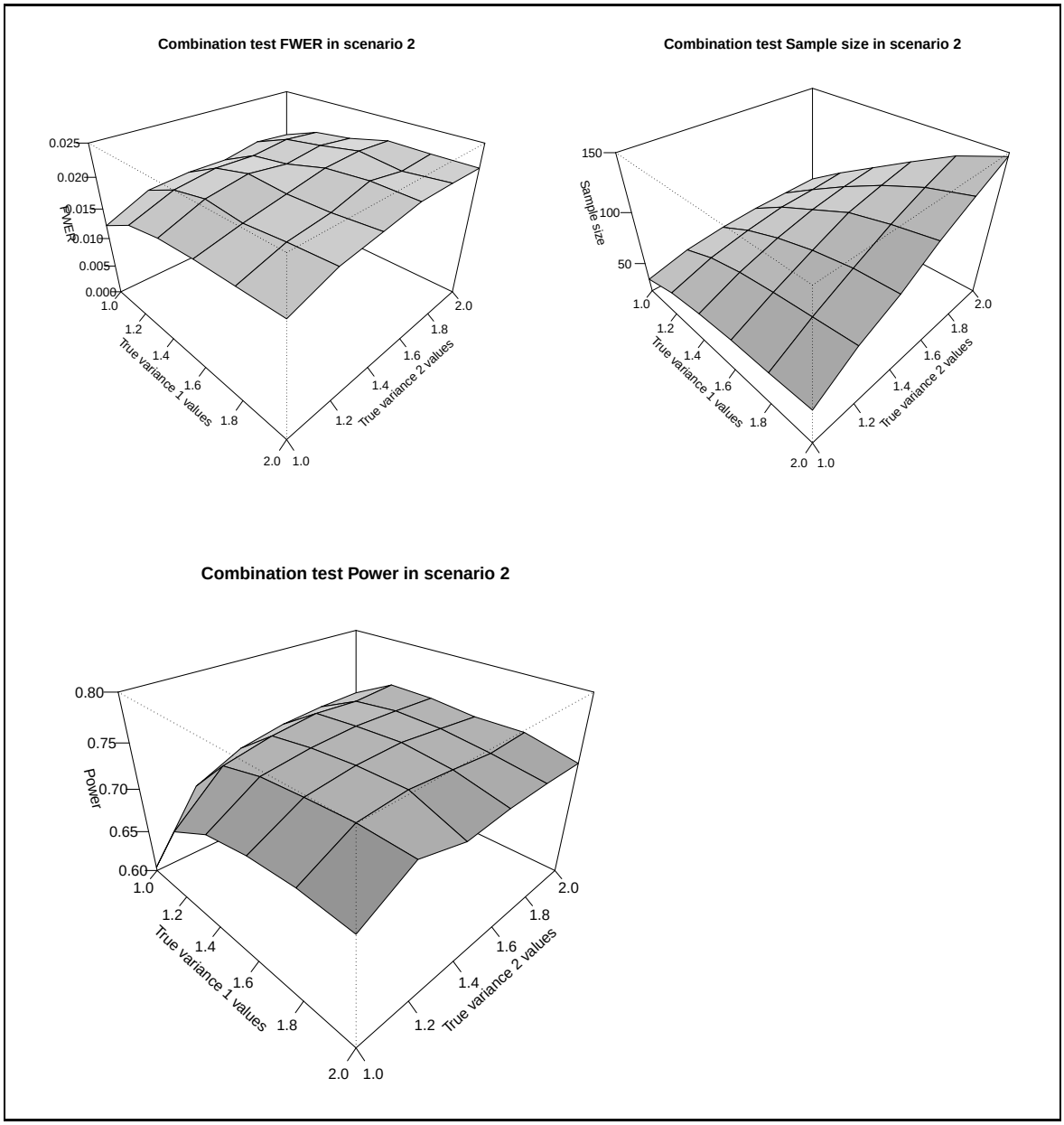

Figure 5.5: GSD combination test FWER, Sample size and Power in Scenario 2

constant values afterwards. The same figure illustrates that the sample size increases in the same direction as σ_1^2 and σ_2^2 with the minimum value 33 and the maximum value 147. Finally the same figure illustrates that the power is not maintained when σ_1^2 and σ_2^2 vary with small values for σ_1^2 = 1.2 and σ_2^2 = 1, and fairly constant values afterwards.

5.6.1.3 Scenario 1 and 2: Summary and comments of the results

The results in scenario 1 shows that the FWER is controlled but becomes increasingly conservative as ρ_{12} increases. We would expect this as the weights in this method are data-dependent. In scenario 2 the results show that the FWER is controlled and fairly constant when ρ_{12} is constant. In all the scenarios, the results obtained show that the FWER is more conservative that in Chapter 4.

The results in scenarios 1 and 2 show that the sample size is increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2, however, the power is decreasing in the opposite direction than the nuisance parameters and this is below that target power. This is a known situation as the weights used in these scenarios do not reflect the sample sizes.

5.6.2 Scenario 3 : Different size effect

5.6.2.1 Scenario 3 : $\delta_1 = 0.5, \delta_2 = 0.7$

Figure 5.6 presents simulation results in Scenario 3 with $\delta_1 = 0.5$ and $\delta_2 = 0.7$. It shows that the FWER is maintained at the nominal 0.025 level despite variation of ρ_{12} and σ_1^2 with the minimum value 0.00957 for perfect correlated data and 0.02088 for uncorrelated data. The same figure shows that the sample size increases as ρ_{12} and σ_1^2 increases with a minimum value of 35 and maximum value of 68. Finally the same figure shows that the power is not maintained but fairly constant when ρ_{12} and σ_1^2 increase with a minimum value of 0.6954 and a maximum value of 0.7178.

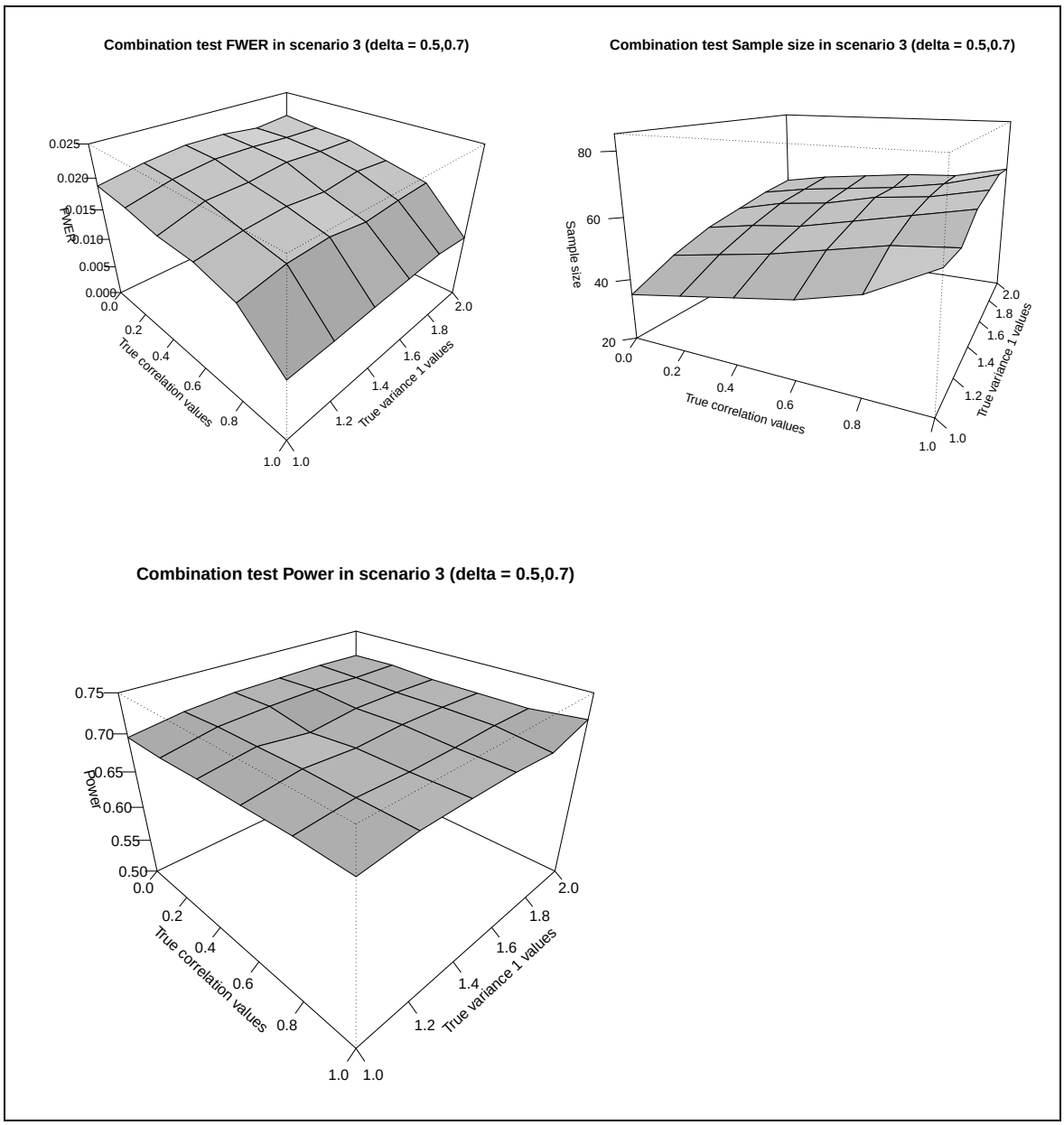

Figure 5.6: GSD combination test FWER, Sample size and Power in Scenario 3 ($\delta_1 = 0.5$, $\delta_2 = 0.7$)

5.6.2.2 Scenario 3: $\delta_1 = 0.7, \delta_2 = 0.5$

Figure 5.7 presents a situation where $\delta_1 = 0.7$ and $\delta_2 = 0.5$. It shows that the FWER in this setting is controlled with a minimum value of 0.00241 for $\rho_{12} = 1$ and a maximum value of 0.02148 for $\rho_{12} = 0$. The same figure shows that the sample size is increasing in the same direction as ρ_{12} and σ_1^2 with a minimum value of 15 and a maximum value of 104. Finally the same figure shows that the power is not maintained and is increasing with σ_1^2 and fairly constant with ρ_{12}, with a minimum value of 0.4002 and a maximum value of 0.7492.

5.6.2.3 Scenario 3: Summary and comments on the results

The results in Scenario 3 show that the FWER is controlled but conservative as ρ_{12} increases. The results in Setting (0.7,0.5) are even more conservative for perfect correlated data.

The results in Scenario 3 also show that sample sizes are increasing in the same direction than ρ_{12} and σ_1^2. However they (sample sizes) are not large enough to detect different effect sizes in both settings at the same time, hence reduction in power. For example, the shape of the power seems to follow the variation of the sample size in Figure 5.7.

5.6.3 Scenario 4: Different spending function

5.6.3.1 Scenario 4: Hwang-Shih-DeCani spending function with $\gamma = -10$

Scenario 4 uses the Hwang-Shih-DeCani spending function with $\gamma = -10$ and Figure 5.8 presents the results with the same input values as in Setting 3. The figure shows that despite variation of ρ_{12} and σ_1^2 the FWER is controlled with a minimum value of 0.01070 and a maximum value of 0.024377. The same figure shows that the sample size increases in the same direction as ρ_{12} and σ_1^2 with the minimum value 57 and the maximum value 125.

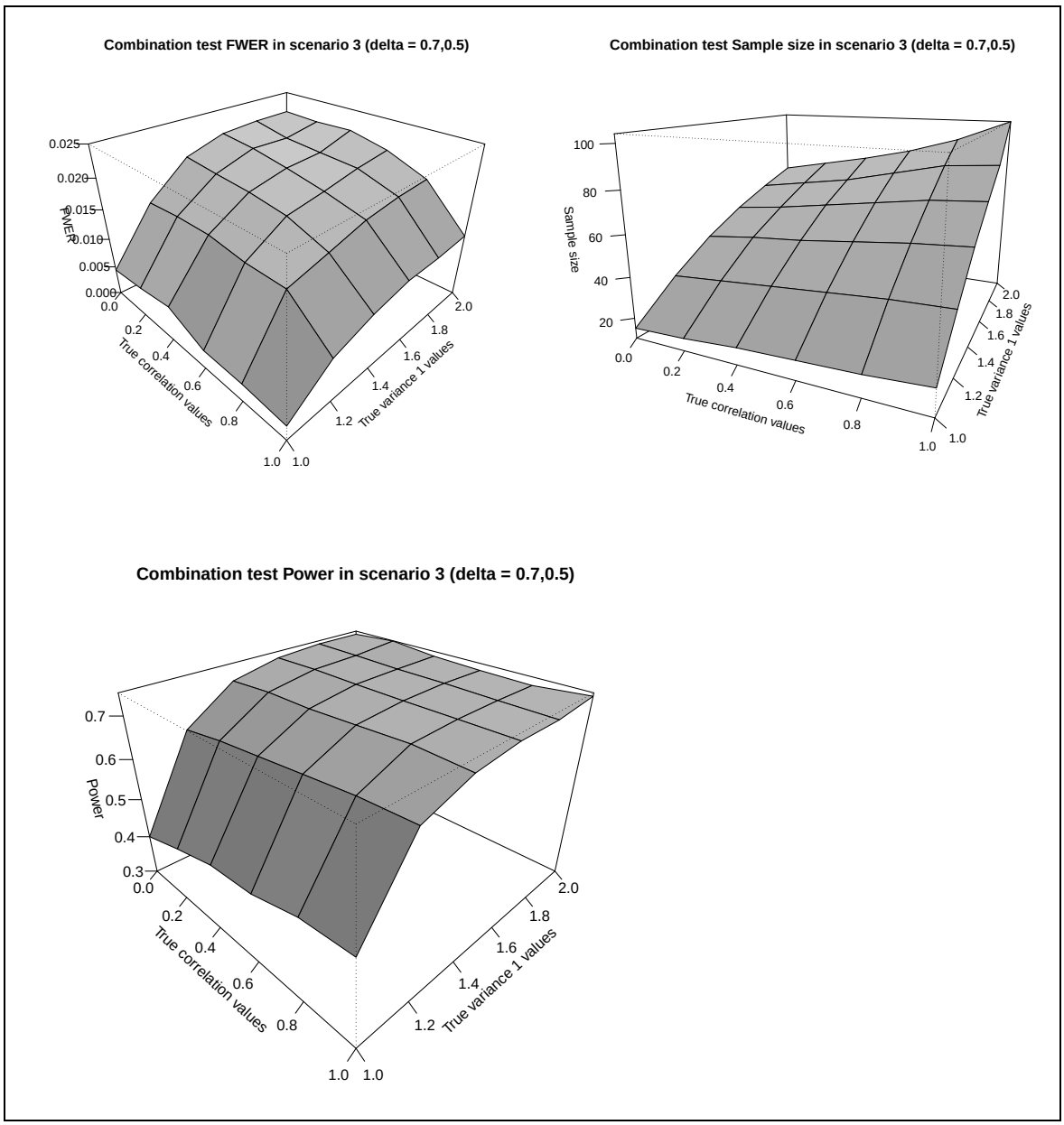

Figure 5.7: GSD combination test FWER, Sample size and Power in Scenario 3 ($\delta_1 = 0.7$, $\delta_2 = 0.5$)

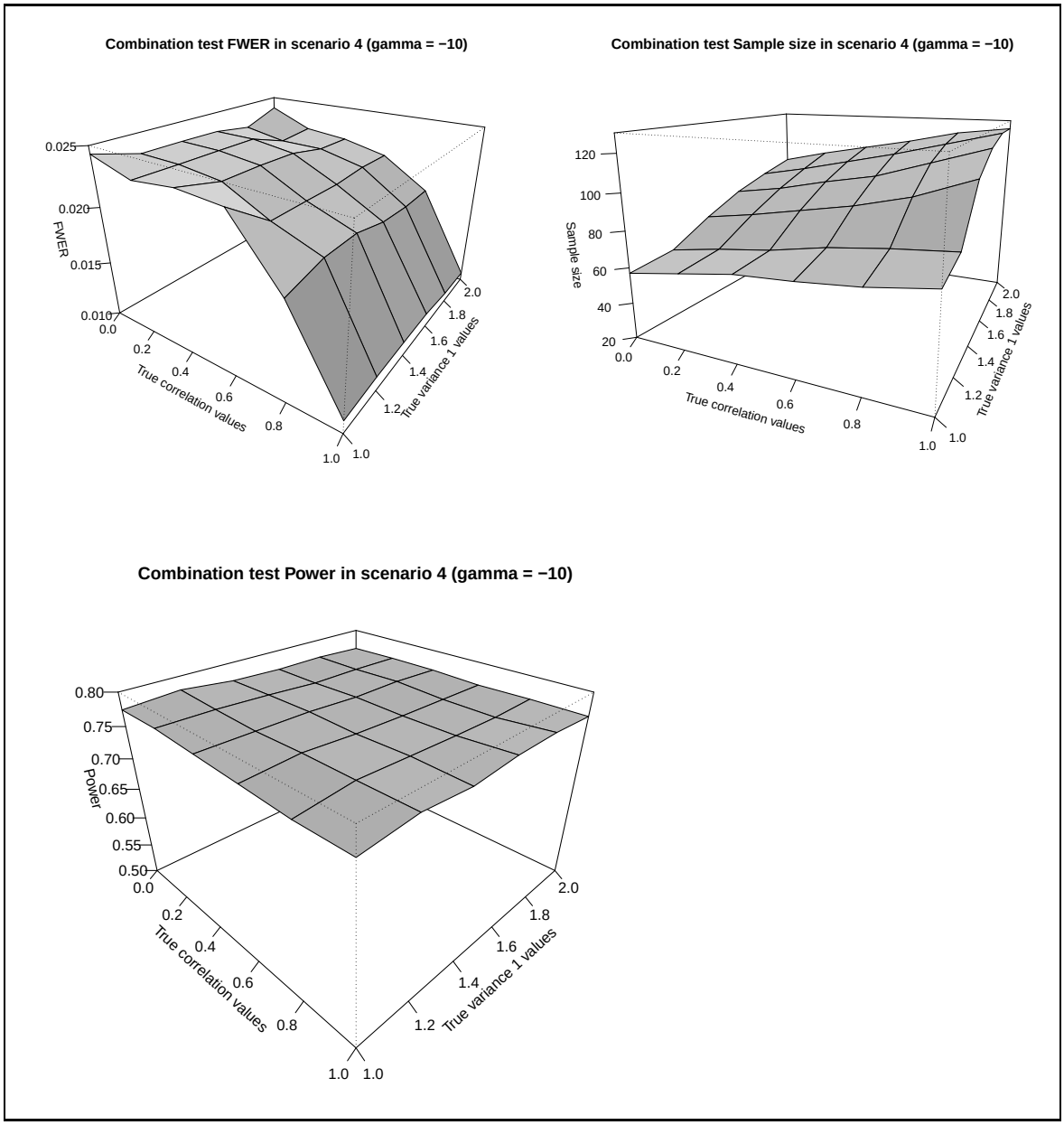

Figure 5.8: GSD combination test FWER, Sample size and Power in Scenario 4 ($\gamma = -10$)

Finally the same figure illustrates that the power is not maintained but fairly constant when ρ_{12} and σ_1^2 increase with a minimum value of 0.7627 and a maximum value of 0.7798.

5.6.3.2 Scenario 4: Hwang-Shih-DeCani spending function with gamma = 10

Scenario 4 also uses the Hwang-Shih-DeCani spending function with $\gamma = 10$. Figure 5.9 presents its results. The figure shows that despite variation of ρ_{12} and σ_1^2, the FWER is controlled with a minimum value of 0.01198 for $\rho_{12} = 1$ and a maximum value of 0.02585 for $\rho_{12} = 0$. The same figure shows that the sample size increases in the same direction as ρ_{12} and σ_1^2 with the minimum value 45 and the maximum value 84. Finally the same figure illustrates that the power is maintained and fairly constant when ρ_{12} and σ_1^2 increase with a minimum value of 0.7840 and a maximum value of 0.8077.

5.6.3.3 Scenario 4: Summary and comments on the results

The results in Scenario 4 are similar to those in Subsection 4.6.1.8. They show that the FWER is controlled but is over conservative as ρ_{12} increases.

They also show that sample sizes are increasing in the same direction as ρ_{12} and σ_1^2. However, the power for $\gamma = 10$ is maintained compared to the one for $\gamma = -10$. This is because Hwang-Shih-DeCani spending function with $\gamma = -10$ gives a very conservative spending function. It spends less at the beginning and more later on as illustrated in Figure 2.1. To reiterate the conclusion of the findings in Subsection 4.6.3.3, by having this type of spending function, it is likely to go to a late look which means there is no effect of the sample size re-estimation because the trial stops at the first look, consequently the power is reduced despite having a big sample size.

Simulation results for this method have shown that in all settings considered, the power is not maintain, except for the setting with the Hwang-Shih-DeCani spending function and $\gamma = 10$. One of the solutions would be to use less conservative spending function

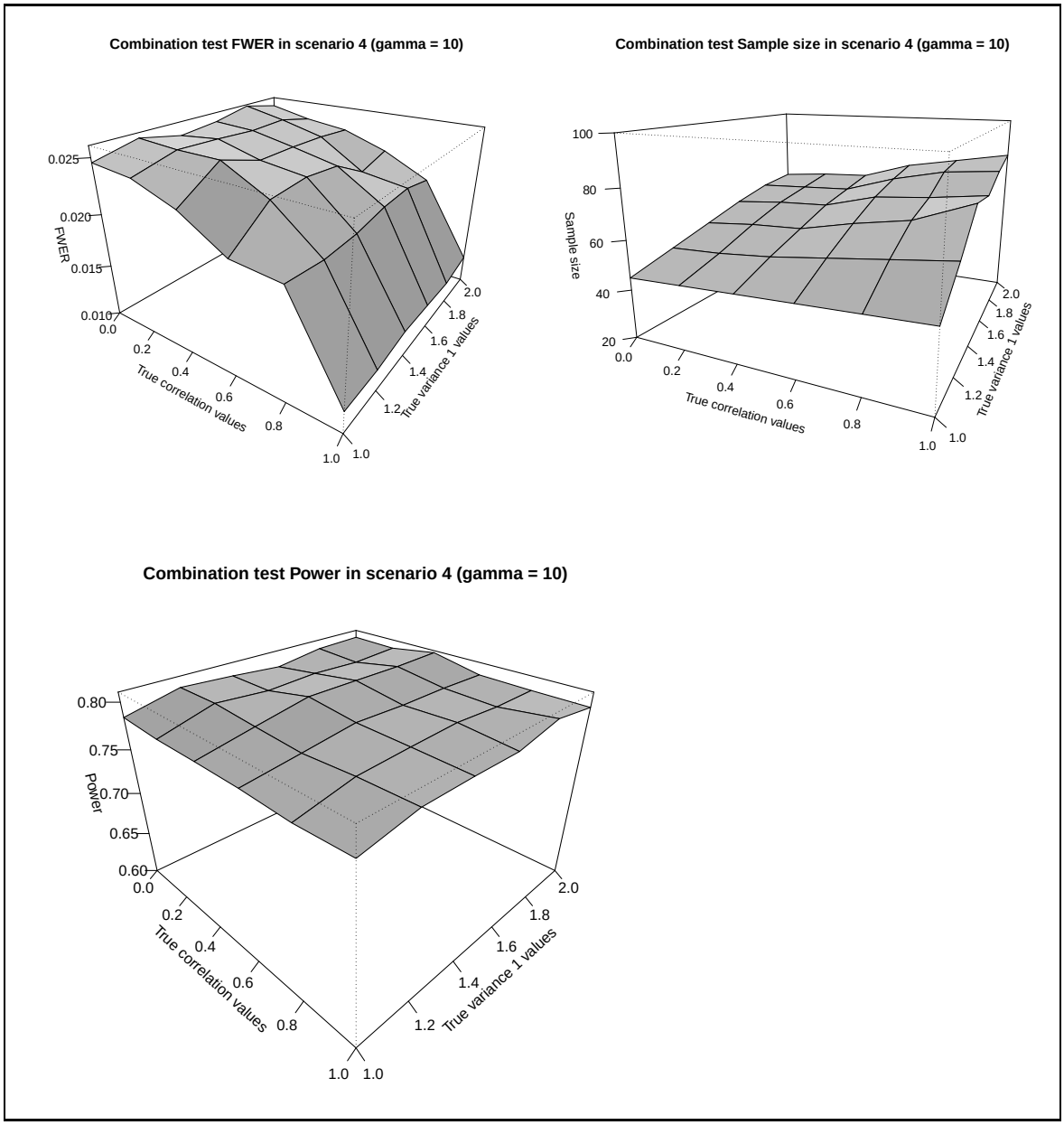

Figure 5.9: GSD combination test FWER, Sample size and Power in Scenario 4 ($\gamma = 10$)

(eg. Hwang-Shih-DeCani spending function with $\gamma = 10$) even if the weights are pre-defined and fixed in advance. Figure 5.9 could be used as an example.

5.7 Summary findings from the simulation results

In this chapter, we described the GSD inverse normal combination tests procedure with multiple co-primary endpoints. The method integrates the concept of an inverse normal combination tests approach into the group sequential designs procedure described in Chapter 4. In Section 5.4, we illustrated how to implement the method in practice and gave an example in Section 5.5. Simulation results presented in Section 5.6 showed that the FWER was controlled but became increasingly conservative as ρ_{12} increased. The results showed no evidence of the inflation of the FWER for all the scenarios considered.

Simulation results presented in Section 5.6 also showed that the sample size was increasing in the same direction as ρ_{12}, σ_1^2 and σ_2^2; however the power was decreasing in the opposite direction than the nuisance parameters and this was below that target power in all settings. We observed similar results as in Chapter 4 regarding the use of different effect sizes simultaneity. To reiterate, we observed that although the sample sizes were increasing in the same direction in Scenario 3, they (sample sizes) were not large enough to detect different effect sizes at the same time, hence reduction in power.

In Section 5.6, we finally showed that the power was not maintain, except for the setting with the Hwang-Shih-DeCani spending function and $\gamma = 10$ (Scenario 4). We then suggested to use this type of setting (Hwang-Shih-DeCani spending function with $\gamma = 10$) if we would like to maintain the power even if the weights were pre-defined and fixed in advance. However, the lack of power in the combination test approach is a known situation because the weights used do not reflect the sample sizes. Authors such as Proschan (2009b) and Lehmacher and Wassmer (1999) have shown this in the setting of a single endpoint.

In the next chapter, we present a discussion and the conclusions of the simulation results of all the methods described in this thesis.

Chapter 6

Discussion and Conclusions

This thesis aimed to show how to adjust a sample size in clinical trials with multiple co-primary continuous endpoints using adaptive and group sequential designs, and also how to construct a test such as to control the FWER and maintain the power, even if the correlation ρ between endpoints is not known. To achieve this, we used three different methods: group sequential designs, sample size re-estimation and inverse normal combination tests.

In this chapter, we present a discussion and the conclusions of the simulation results of the three methods previously described. Section 6.1 presents a discussion of the results, Section 6.2 presents extensions and further work and Section 6.3 presents the conclusions.

6.1 Discussion

6.1.1 Sample size re-estimation method

The sample size re-estimation method was introduced in the context of a single endpoint in Section 2.1, then extended in the context of multiple co-primary endpoints in Section 3.1. We used the blinded method, and then considered the case of two co-primary endpoints where the rejection of either is considered as proof that the new treatment is working. We conducted two different tests each for one endpoint and at level α/K, and implemented the

Internal Pilot Study design as developed by Wittes and Brittain (1990). The new features for this method are: use of the Bonferroni correction to adjust for the FWER, steps to follow when designing a clinical trial with multiple co-primary endpoints, steps to follow when calculating sample size with multiple co-primary endpoints, what to do if different effect sizes are needed in a clinical trial with multiple endpoints. Simulation results in Section 3.1.7 show that this method controls the FWER and maintains the power; and these results are similar to the results obtained by Friede and Schmidli (2010). However, it is important stress that the control of the FWER is not guaranteed analytically but appears to hold in the SSR case. We also show that the power could not be maintained if strange revision rules are used. For example, the interim evaluation performed when 10% of the patients are in the internal pilot leads to inaccurate estimation of the nuisance parameters, with the consequence that the power was not maintained. This also applies when different effect sizes are used simultaneously.

6.1.2 SSR Inverse Normal Combination test method

The SSR Inverse Normal Combination test method was presented as a method integrating the concept of the inverse normal combination test into sample size re-estimation. This means that the same design as proposed by Wittes and Brittain (1990) was used to perform sample size re-estimation, but the only difference was how the final analysis was conducted. It was first introduced in the context of single endpoint in Section 2.2, followed by its extension in the context of multiple co-primary endpoints in Section 3.2. Here, we used the inverse normal combination test method as proposed by Lehmacher and Wassmer (1999), to construct the test statistics. The new features for this method are: use of the Bonferroni correction to adjust for the FWER, steps to follow when designing a clinical trial with multiple co-primary endpoints and the application of the method in the context of the internal pilot study design. Simulation results in Subsection 3.2.6 show that this method controls

the FWER, but there is a cost, which is a loss of power, because the weights of the combination test are not based on the observed sample size but on weights fixed in advance. Although we would expect this method not to maintain the power for the reasons explained earlier, simulation results show that by changing the timing of the interim intervention, this method maintains the power even if it uses pre-defined weights fixed in advance. Simulation (analytically too) results also show that changing the weights allocation has an impact on maintaining the power, however, finding the optimal weights allocation is challenging because they (weights) are defined before the trial begins.

6.1.3 Group Sequential Designs method

We first introduced the group sequential method in the context of a single endpoint in Section 2.3, as described in more detail by Jennison and Turnbull (2000a). We then extended it in the context of multiple co-primary endpoints in Chapter 4, where K group sequential tests were conducted, each for one endpoint and at level α/K. We considered the case of K co-primary endpoints where the rejection of any is taken into account as proof that the new treatment is conclusive. We designed the method in such a way to allow for early stopping, at the same time to monitor the information, which was adjusted at each stage to allow for estimating the variance σ^2. The solution that we propose to solve the problem defined in Subsection 1.4.2, is an easy and simple option, it can be implemented in phase III of a clinical trial without difficulties. It is also totally different to the solutions of Jennison and Turnbull (1993) and Cook and Farewell (1994). The method proposed is flexible in such a way that it allows a group sequential design planned with *J* stages to be modified to a group sequential design with *J* + 1 stages. This is something new that neither Jennison and Turnbull (1993) nor Cook and Farewell (1994) proposed. Simulation results showed that the FWER was controlled but became increasingly conservative as ρ_{12} for large correlated endpoints. The results also showed that the FWER was above the nominal level of 0.025

for uncorrelated data, i.e. ρ_{12} = 0. Nevertheless, this slight increase in the FWER was less than 0.001, hence too small to be practically relevant. Therefore we concluded that for the scenarios considered here, the GSD procedure controls the FWER. Simulation results also show that, although the method maintains the power reasonably well, it may lack power depending on the spending function used. For example we showed that the power was not maintained for the Hwang-Shih-DeCani spending function with $\gamma = -10$.

6.1.4 GSD Inverse Normal Combination test method

The GSD Inverse Normal Combination test method was presented as a method which integrates the concept of inverse normal combination tests illustrated in Section 2.4 into GSD as described in Chapter 4. It was first introduced in the context of single endpoint in Section 2.2 and Section 2.4, followed by its extension in the context of multiple co-primary endpoints in Chapter 5. The new features for this method are: use of the Bonferroni correction to adjust for the FWER, steps to follow when designing a clinical trial with multiple co-primary endpoints and the application of the method in the context of GSD. Simulation results in Section 5.6 show that this method controls the FWER, but there is a cost which is a loss of power; however, the results show that one of the solutions for this would be to use a less conservative spending function (eg. Hwang-Shih-DeCani spending function with $\gamma = 10$) even if the weights are pre-defined and fixed in advance. It is important to highlight that the FWER of this method becomes increasingly conservative as the correlation between endpoints increases compared to the GSD method.

6.2 Extensions and future work

We presented different solutions for sample size re-estimation in phase III of a clinical trial with multiple co-primary endpoints. We gave simulation examples for two continuous

endpoints. The first possible extension for all the methods would be to consider an example of more than two endpoints and investigate whether the FWER would still be controlled and the power still maintained. The second extension would be to consider two sided-test scenarios and check how this would affect the FWER and the power.

In Chapter 4, further work could be done on the GSD method with multiple co-primary endpoints by adding lower boundaries into the design. Suppose c_{ju} and c_{jl} represent the upper boundary and lower boundary, respectively. The stopping rules defined in Eq. (2.39) are extended as follows:

$$\begin{array}{ll} \text{After group j = 1, ..., J-1} & \\ \quad \text{if } Z_j \geq c_{ju} & \text{stop, reject } H_0 \\ \quad \text{if } Z_j < c_{jl} & \text{stop, do not reject } H_0 \\ \quad \text{otherwise} & \text{continue to group j +1} \\ \text{after group J} & \\ \quad \text{if } Z_J \geq c_{Ju} & \text{stop, reject } H_0 \\ \quad \text{if } Z_J < c_{Jl} & \text{stop, do not reject } H_0 \\ \quad \text{otherwise} & \text{stop, accept } H_0 \end{array} \tag{6.1}$$

The type I error is defined through the critical values $c_{1u}, ..., c_{Ju}$ and $c_{1l}, ..., c_{Jl}$ as described in Eq. (6.2) below for a single endpoint:

$$Pr(c_{1l} < Z_1 < c_{1u}, ..., c_{(j-1)l} < Z_{j-1} < c_{(j-1)u}, c_{jl} > Z_j > c_{ju} | \theta = 0) = \pi_j / K \tag{6.2}$$

calculated when the $Z_1, ..., Z_J$ follow the distribution of Eq. (2.38) and the stopping rules of Eq. (6.1). Considering an example of two endpoints and j stages, the extension of the FWER defined in Subsection 4.2.3 is now:

At stage 1, we have:

Pr(stop and reject at least one H_{0k} at or before stage 1 | $\theta_k = 0$) =

$P(Z_{11} > c_{1u}$ or $Z_{21} > c_{1u}$ | θ=0) + $P(Z_{11} < c_{1l}$ and $Z_{21} < c_{1l}$ | θ=0) $\leq \pi_1$

At look 2, we have:

Pr(stop and reject at least one H_{0k} at or before stage 2 | $\theta_k = 0$) =

$Pr(Z_{11} > c_{1u}$ or $Z_{21} > c_{1u}$ | θ=0) + $Pr(Z_{11} < c_{1l}$ and $Z_{21} < c_{1l}$ | θ=0) +

$Pr(c_{1l} < Z_{11} < c_{1u}$, $c_{1l} < Z_{21} < c_{1u}$,$Z_{12} > c_{2u}$ or $Z_{22} > c_{2u}$, $Z_{12} < c_{2l}$

and $Z_{22} < c_{2l}$,| $\theta = 0) \leq \pi_2$

At look j, we have:

Pr(stop and reject at least one H_{0k} at or before stage j | $\theta_k = 0$) =

$Pr(Z_{11} > c_{1u}$ or $Z_{21} > c_{1u}$ | θ=0) + $Pr(Z_{11} < c_{1l}$ and $Z_{21} < c_{1l}$ | θ=0) +

$Pr(c_{1l} < Z_{11} < c_{1u}$, $c_{1l} < Z_{21} < c_{1u}$,$Z_{12} > c_{2u}$ or $Z_{22} > c_{2u}$, $Z_{12} < c_{2l}$

and $Z_{22} < c_{2l}$,| $\theta = 0) + ... + Pr(c_{1l} < Z_{11} < c_{1u}$, $c_{1l} < Z_{21} < c_{1u}$,

$c_{2l} < Z_{12} < c_{2u}$,$c_{2l} < Z_{22} < c_{2u}$,...,$c_{(j-1)l} < Z_{1(j-1)} < c_{(j-1)u}$,

$c_{(j-1)l} < Z_{2(j-1)} < c_{(j-1)u}$, $Z_{1j} > c_{ju}$ or $Z_{2j} > c_{ju}$, $Z_{1j} < c_{jl}$ and

$Z_{2j} < c_{jl}$ | $\theta = 0) \leq \pi_j$ (6.3)

π_j represents the error spending function at stage j, j = 1,...,J. So, to control the FWER, one must choose c_{ju} and c_{jl} to satisfy Eq. (6.2) (i.e, calculate c_{ju} and c_{jl} using Eq. (6.2), and workout π_j).

The power defined in Eq. (4.11) is extended, so now we have:

$$
\begin{aligned}
&Pr(\text{stop and reject at least one } H_{0k} \text{ at or before stage j} \mid \theta_k = \delta) = \\
&Pr(Z_{11} > c_{1u} \text{ or } Z_{21} > c_{1u} \mid \theta = \delta) + Pr(Z_{11} < c_{1l} \text{ and } Z_{21} < c_{1l} \mid \theta = \delta) + \\
&Pr(c_{1l} < Z_{11} < c_{1u}, c_{1l} < Z_{21} < c_{1u}, Z_{12} > c_{2u} \text{ or } Z_{22} > c_{2u}, Z_{12} < c_{2l} \\
&\text{and } Z_{22} < c_{2l}, \mid \theta = \delta) + ... + Pr(c_{1l} < Z_{11} < c_{1u}, c_{1l} < Z_{21} < c_{1u}, \\
&c_{2l} < Z_{12} < c_{2u}, c_{2l} < Z_{22} < c_{2u}, ..., c_{(j-1)l} < Z_{1(j-1)} < c_{(j-1)u}, \\
&c_{(j-1)l} < Z_{2(j-1)} < c_{(j-1)u}, Z_{1j} > c_{ju} \text{ or } Z_{2j} > c_{ju}, Z_{1j} < c_{jl} \text{ and} \\
&Z_{2j} < c_{jl} \mid \theta = \delta) = 1 - \beta
\end{aligned}
\tag{6.4}
$$

For given values of j, α, β, c_{ju} and c_{jl}, the maximum sample size can be found using mvtnorm package in R to compute multivariate normal probabilities (see Subsection 3.1.4).

6.3 Conclusions

As a conclusion, in this thesis we consider statistical methods for dealing with interim data in phase III of a clinical trial with multiple co-primary endpoints. They are: sample size re-estimation, the group sequential approach and the inverse normal combination test procedure. We show that all the methods control the FWER and we explain in which settings we would expect the inflation of the FWER. We also show that the SSR and GSD methods maintain the power and that the GSD seems to be more powerful, again highlighting settings where the power is not maintained. We finally show that the power of the combination test method is not maintained and explain the reasons for this. The conclusion about the findings imply that what we can do with the GSD, we also can do with SSR and combination test method. Simulations presented in this thesis have shown that even if the results are more or less similar, we can do it in different ways, and the GSD has proven to

be more powerful. The solutions we propose are either new methods to problems that have not yet been solved, or approaches that are simpler to apply, and more efficient than ones currently existing in the literature. We recommend our approaches (SSR and GSD) as they are simpler to apply than existing ones.

Appendix A: SSR simulation program

```
#############################################################################
## This program contains the mean vector, the variance covariance matrix
## and the multivariate normal probability function.  The program computes
## the maximum sample size for SSR method and performs the test of hypothesis
## and append the results at the end.
#############################################################################

## Load the pachages below if needed

library(lattice)
library(ldbounds)
library(mvtnorm)

## Defining the function containing the non-centrality parameter, the
## variance covariance matrix and the multivariate normal probability function

sen <- function(n,stages,rho,diff,s1,s2){
```

```
### variance covariance for look2 ##
sigma3 <- matrix(c(1,rho,
rho,1),
nrow=2,ncol=2)

#### Noncentrality parameter for look 2 ####

mean12 <-  (diff/s1)*sqrt((2*n)/2)
mean22 <-  (diff/s2)*sqrt((2*n)/2)

###DEFINING FUCNTION####

out <- 1 - pmvnorm(lower=c(-Inf,-Inf), upper=c(boundary[1],boundary[1]),
mean=as.numeric(c(mean12,mean22)),sigma=sigma3)-0.8
}

##### Set up the following parameters ######################

rho <- 0.5 # correlation betwen endpoints
s1 <- 1.5  # stadard deviation for endpoint 1
s2 <- 1    # stadard deviation for endpoint 1
diff <- 0.5 # clinically significant treatment difference
stages <- 2  # number of stage
```

```
alpha <- 0.0125 # nominal alpha
mu1 <- matrix(c(diff,diff),nrow=1,ncol=2) # mean vector
mu0 <- matrix(c(0,0),nrow=1,ncol=2) # mean vector

## Calculate the initial maximum sample size
n1 <- floor(uniroot(sen,lower=1,upper=1000,stages=stages,rho=rho,
    diff=diff,s1=s1,s2=s2)$root)
dfe1 <- (2*n1)-2 # degree of freedom

# simulation set-up
nsim <- 100000
set.seed(1)

#### set up a file where  sigmas and maximum sample size and number
# of rejection are stored
results <- matrix(0,nrow=nsim,ncol=11)
colnames(results) <- c("Gesrho", "GesS1", "GesS2", "n1", "Trurho",
"Estrho", "EstS1", "EstS2", "TotalN",  "efficacy")

# run simulations
for (i in 1:nsim){

results[i,1] <- rep(rho) # append guessed correlation
results[i,2] <- rep(s1) # append guessed sigma 1 for endpoint 1
```

```
results[i,3] <- rep(s2) # append guessed sigma 2 for endpoint 2
results[i,4] <- rep(2*n1) # append initial sample size

# initiate
istop <- 0

# simulate sample of size n1
Trurho <- 0 # true correlation
Trus1 <- 1 # true sigma 1
Trus2 <- 1 # true sigma 2
results[i,5] <- rep(Trurho) # append  true correlation
nsigma <- matrix(c(Trus1^2,Trus1*Trus2*Trurho,Trus1*Trus2*Trurho,Trus2^2),
ncol=2,nrow=2) # calculate variance covariance matrix
sample <- rmvnorm(n1,c(0,0),nsigma) # simulate sample for treatment group
Csample <- rmvnorm(n1,mu0,nsigma) # simulate sample for control group
nrho <- cor(c(Tsample[,1],Csample[,1]),c(Tsample[,2],Csample[,2]))
    # estimate correlation
results[i,6] <- nrho # append estimated correlation
s11 <- sqrt(var(c(Tsample[,1],Csample[,1]))) # estimate variance for endpoint 1
results[i,7] <- s11 # append estimated variance 1
s22 <- sqrt(var(c(Tsample[,2],Csample[,2]))) # estimate variance for endpoint 2
results[i,8] <- s22 # append estimated variance 2

## Re-estimated maximum sample size
```

```
n2 <- round(uniroot(sen,lower=1,upper=1000,stages=stages,rho=nrho,
    diff=diff,s1=s11,s2=s22)$root)
N <- 2*n2 # total re-estimated maximum sample szie
df <- (2*N)-2 # calculate degree of freedom
results[i,9] <- N # append maximum sample size

if (istop==0) {
if (n1>=N) {istop <- 1}
else {

# simulate sample of size N-n1
Tsamplev <- rmvnorm(N-n1,c(0,0),nsigma) # simulate sample of size N-n1 for
     treatment group
Csamplev <- rmvnorm(N-n1,mu0,nsigma) # simulate sample of size N-n1 for
    control group
Tsample <- rbind(Tsample,Tsamplev) # combine stage 1 data and stage 2 data
     for treatment group
Csample <- rbind(Csample,Csamplev) # combine stage 1 data and stage 2 data
     for control group

s1f <- sqrt(var(c(Tsample[,1],Csample[,1]))) # estimated variance 1 for
    combined data for endpoint 1
s2f <- sqrt(var(c(Tsample[,2],Csample[,2]))) # estimated variance 1 for
    combined data for endpoint 2
```

```
cva <- qt(1-alpha,df) # critical value

# test statistics
z1 <- (mean(Tsample[,1])-mean(Csample[,1]))/(s1f*sqrt(2/N)) # calculate
    t test for endpoint 1
z2 <- (mean(Tsample[,2])-mean(Csample[,2]))/(s2f*sqrt(2/N)) # calculate
    t test for endpoint 2

efficacy <- as.numeric(z1>=(cva) | z2>=(cva)) # number of rejection H0
results[i,10] <- efficacy # append number of rejections

### end of final analysis ###
 }# end if
} # end simulation
}
# append results
    write.table(results,file="./Ztestguess05true00null.txt",
    sep="\t",eol="\n",col.names=TRUE,na = "NA",row.names=FALSE)
```

Appendix B: SSR inverse normal combination test simulation program

```
##############################################################################
## This program contains the mean vector, the variance covariance matrix
## and the multivariate normal probability function.  The program computes
## the maximum sample size for Inverse Normal Combination test method and
## performs the test of hypothesis and append the results at the end.
##############################################################################

## Load the pachages below if needed

library(lattice)
library(ldbounds)
library(mvtnorm)
## Defining the function containing the non-centrality parameter, the
## variance covariance matrix and the multivariate normal probability function
```

```
sen <- function(n,stages,rho,diff,s1,s2){

### variance covariance for look2 ##
sigma3 <- matrix(c(1,rho,
rho,1),
nrow=2,ncol=2)

#### Noncentrality parameter for look 2 ####

mean12 <-  (diff/s1)*sqrt((2*n)/2)
mean22 <-  (diff/s2)*sqrt((2*n)/2)

###DEFINING FUCNTION####

out <- 1 - pmvnorm(lower=c(-Inf,-Inf), upper=c(boundary[1],boundary[1]),
mean=as.numeric(c(mean12,mean22)),sigma=sigma3)-0.8
}

##### Set up the following parameters ######################

rho <- 0.5 # correlation betwen endpoints
s1 <- 1.5  # stadard deviation for endpoint 1
s2 <- 1    # stadard deviation for endpoint 1
diff <- 0.5 # clinically significant treatment difference
```

```
stages <- 2  # number of stage
alpha <- 0.0125 # nominal alpha
mu1 <- matrix(c(diff,diff),nrow=1,ncol=2) # mean vector
mu0 <- matrix(c(0,0),nrow=1,ncol=2) # mean vector

## Calculate the initial maximum sample size
n1 <- floor(uniroot(sen,lower=1,upper=1000,stages=stages,rho=rho,diff=diff,
    s1=s1,s2=s2)$root)
dfe1 <- (2*n1)-2 # stage 1 degree of freedom

# simulation set-up
nsim <- 100000
set.seed(1)

#### set up a file where  sigmas and maximum sample size and number
# of rejection are stored
results <- matrix(0,nrow=nsim,ncol=11)
colnames(results) <- c("Gesrho", "GesS1", "GesS2", "n1", "Trurho",
"Estrho", "EstS1", "EstS2", "TotalN",  "efficacy")

# run simulations
for (i in 1:nsim){

results[i,1] <- rep(rho) # append guessed correlation
```

```
results[i,2] <- rep(s1) # append guessed sigma 1 for endpoint 1
results[i,3] <- rep(s2) # append guessed sigma 2 for endpoint 2
results[i,4] <- rep(2*n1) # append initial sample size

# initiate
istop <- 0

# simulate sample of size n1
Trurho <- 0 # true correlation
Trus1 <- 1 # true sigma 1
Trus2 <- 1 # true sigma 2
results[i,5] <- rep(Trurho) # append  true correlation
nsigma <- matrix(c(Trus1^2,Trus1*Trus2*Trurho,Trus1*Trus2*Trurho,Trus2^2),
ncol=2,nrow=2) # calculate variance covariance matrix
sample <- rmvnorm(n1,c(0,0),nsigma) # simulate sample for treatment group
Csample <- rmvnorm(n1,mu0,nsigma) # simulate sample for control group
nrho <- cor(c(Tsample[,1],Csample[,1]),c(Tsample[,2],Csample[,2]))
    # estimate correlation
results[i,6] <- nrho # append estimated correlation
s11 <- sqrt(var(c(Tsample[,1],Csample[,1]))) # estimate variance for
    stage 1 data endpoint 1
results[i,7] <- s11 # append estimated variance 1
s22 <- sqrt(var(c(Tsample[,2],Csample[,2]))) # estimate variance for
    stage 1 endpoint 2
```

```
results[i,8] <- s22 # append estimated variance 2

######### Combinatiion test stage 1 ##############

t1 <- (mean(Tsample[,1])-mean(Csample[,1]))/(s11*sqrt((1/n1)+(1/n1)))
    # t test for endpoint 1
t2 <- (mean(Tsample[,2])-mean(Csample[,2]))/(s22*sqrt((1/n1)+(1/n1)))
    # t test for endpoint 2
p1 <- (1-pt(t1,dfe1)) # p value for endpoint 1
p2 <- (1-pt(t2,dfe1)) # p value for endpoint 2

## Inverse Normal Combination
testcum11=(sqrt(0.5))*qnorm(1-p1) # stage 1 test for endpoint 1
testcum21=(sqrt(0.5))*qnorm(1-p2) # stage 1 test for endpoint 2

### End combination test#####

######### Combination test stage 2 ##############
    n2 <- round(uniroot(sen,lower=1,upper=1000,stages=stages,
    rho=nrho,diff=diff,s1=s11,s2=s22)$root) # maximum sample size re-estimated
    N <- 2*n2 # total maximum sample size
    dfe2 <- (2*(N-n1))-2 # stage 2 degree of freedom

    results[i,9] <- N # append  maximum sample size
```

```
if (istop==0) {

if (n1>=N) {istop <- 1}

else {
# simulate sample of size N-n1
Tsamplev <- rmvnorm(N-n1,c(0,0),nsigma) # simulate sample of size N-n1 for
     treatment group
Csamplev <- rmvnorm(N-n1,mu0,nsigma) # simulate sample of size N-n1 for
    control group
s111 <- sqrt(var(c(Tsamplev[,1],Csamplev[,1]))) # estimate variance for
    stage 2 endpoint 1
s222 <- sqrt(var(c(Tsamplev[,2],Csamplev[,2]))) # estimate variance for
    stage 2 endpoint 2

# test statistics
t11 <- (mean(Tsamplev[,1])-mean(Csamplev[,1]))/(s111*sqrt((1/(N-n1))+
    (1/(N-n1)))) # t test for endpoint 1
t22 <- (mean(Tsamplev[,2])-mean(Csamplev[,2]))/(s222*sqrt((1/(N-n1))+
    (1/(N-n1)))) # t test for endpoint 2
p11 <- (1-pt(t11,dfe2))  # p value for endpoint 1
p22 <- (1-pt(t22,dfe2)) # p value for endpoint 2
```

```
    testcum12=(sqrt(0.5))*qnorm(1-p11) # stage 2 test for endpoint 1
    testcum22=(sqrt(0.5))*qnorm(1-p22) # stage 2 test for endpoint 2

################     End combination test   #############
#########     Final Analysis   #############
    test1 = testcum11 + testcum12 # combined test for stage 1 and stage 2 data
    endpoint 1
    test2 = testcum21 + testcum22 # combined test for stage 1 and stage 2 data
    endpoint 2
    efficacy <- as.numeric(test1>=(boundary[1]) | test2>=(boundary[1]))
    # number for accepting or rejection H0
    results[i,10] <- efficacy # append number of rejections

 }# end if
} # end simulation
}

# append results
    write.table(results,file="./Ztestguess05true02null.txt",sep="\t",eol="\n",
    col.names=TRUE,na = "NA",row.names=FALSE)
```

Appendix C: Program to compute the boundaries of a GSD

```
##############################################################################
#This program contains the main program to compute the boundaries  using
Simpson method
##############################################################################

simpson<-function(f,dx){

# Approximates integrals using Simpsons rule.
# f is a vector of function values at 2m+1 equally spaced x
# values (comprising 2m intervals of length dx).

m<-(length(f)-1)/2
evens<-2*(1:m);
odds<-2*(1:m)+1;
last<-2*m+1
```

```
int<-(4*sum(f[evens])+2*sum(f[odds])+f[1]-f[last])*(dx/3) return(int) }


updt<-function(datavec){


# updt gives P(Z_1\le c_1,...,ldots,Z_prev\le c_prev,
# Zcur=zcur);
# datavec consists of:
# First dim1 components contain fprev=P(Z_1\le c_1,\ldots,
# Z_{prev-1}\le c_{prev-1},Z_prev=zprev) for a vector  prevgrid.
# Next dim1 components contain the vector prevgrid.
# Next 2 components contain information times of previous and current looks.
# The last component contains zcur.


dimmy<-length(datavec);
dim1<-(dimmy-3)/2;
dim1p<-dim1+1
dim2<-2*dim1
dim2p<-dim2+1;
dim2pp<-dim2+2;
dim2ppp<-dim2+3
fprev<-datavec[1:dim1];
prevgrid<-datavec[dim1p:dim2]
tprev<-datavec[dim2p]
tcur<-datavec[dim2pp];
```

```
zcur<-datavec[dim2ppp]
temp<-(zcur*sqrt(tcur)-prevgrid*sqrt(tprev))/sqrt(tcur-tprev)
y<-sqrt(tcur/(tcur-tprev))*exp(-temp^2/2)/sqrt(2*pi)*fprev
dx<-prevgrid[2]-prevgrid[1] return(simpson(y,dx)) }

distrib<-function(tt,cc){

# Gives P(Z(t_1)\le c_1,...Z(t_k)\le c_k), where
# cc=(c_1,\ldots,c_k),
# tt=(t_1,\ldots,t_k)

if(length(tt)!=length(cc)){return("dimensions of t and c do not match")}
if(max(tt)>1 | min(tt)<=0){return("t_i not in (0,1] for all i")}
if(min(cc)<-7){return(0)}
k<-length(tt)
numint<-50
lengthz<-numint+1
if(k==1){return(pnorm(cc[1],0,1))}
zgrid<-seq(-7,cc[1],length=lengthz)
tprev<-tt[1]

fprev<-exp(-zgrid^2/2)/sqrt(2*pi) for(i in 2:k){
zcur<-seq(-7,cc[i],length=lengthz)
tprev<-tt[i-1]
```

```
tcur<-tt[i]
datmat<-matrix(rep(c(fprev,zgrid,tprev,tcur),lengthz), nrow=lengthz, byrow=T)
datmat<-cbind(datmat,zcur)
ww<-apply(datmat,MARGIN=1,updt)
fprev<-ww
zgrid<-zcur }

dz<-zcur[2]-zcur[1]

ans<-simpson(ww,dz) return(ans) }

################################################################
rename<-function(cccur,otherstuff){

# Re-parameterizes the distribution function P(Z(t_1)\le c_1,
# \ldots,Z(t_k)\le c_k) so that first variable is c_k and the
# last variable is the cumulative alpha spent.

kminus1<-(length(otherstuff)-2)/2
i1<-kminus1+1;
i2<-kminus1+2;
i3<-2*kminus1+1;
i4<-2*kminus1+2
ttcur<-otherstuff[1];
```

```
ttprev<-otherstuff[2:i1]
ccprev<-otherstuff[i2:i3]
alphacum<-otherstuff[i4]
return(distrib(c(ttprev,ttcur),c(ccprev,cccur))-(1-alphacum)) }

findroot<-function(f,low,high,otherstuff){

# Finds root in interval (low,high) for the increasing (in x) pending Functions
# function f(x,otherstuff).

lower<-low;
higher<-high
if(f(lower,otherstuff)>0 | f(higher,otherstuff)<0){return
("findroot find the root")} for(i in 1:20){
midpoint<-(lower+higher)/2
if(f(midpoint,otherstuff) > 0){higher<-midpoint}
else{lower<-midpoint} }
return((lower+higher)/2)}
```

Appendix D: Program containing the mean vector, the covariance matrix and the multivariate probability function

```
#############################################################################
## This program contains the mean vector, the variance covariance matrix
## and the multivariate normal probability function
#############################################################################

gs2.power <- function(n,stages,rho,diff,bound,s1,s2){

################################### set-up stuff ############################
 if(stages<2 | stages>5){
  stop("Stages must be <=5 and >1")
 } # end if
 b <- bound
```

```
##### Noncentrality parameter for look 1 ####
mean11 <-  (diff/s1)*sqrt(n/2)
mean21 <-  (diff/s2)*sqrt(n/2)

#### Noncentrality parameter for look 2 ####
mean12 <-  (diff/s1)*sqrt(2*n/2)
mean22 <-  (diff/s2)*sqrt(2*n/2)

#### Noncentrality parameter for look 3 ####
mean13 <-  (diff/s1)*sqrt((3*n)/2)
mean23 <-  (diff/s2)*sqrt((3*n)/2)

### variance covariance for look3 ##
sigma3 <- matrix(c(1,rho,sqrt(1/2),sqrt(1/2)*rho,sqrt(1/3),sqrt(1/3)*rho,
rho,1,sqrt(1/2)*rho,sqrt(1/2),sqrt(1/3)*rho,sqrt(1/3),
sqrt(1/2),sqrt(1/2)*rho,1,rho, sqrt(2/3),sqrt(2/3)*rho,
sqrt(1/2)*rho,sqrt(1/2),rho,1,sqrt(2/3)*rho,sqrt(2/3),
sqrt(1/3),sqrt(1/3)*rho,sqrt(2/3),sqrt(2/3)*rho,1,rho,
sqrt(1/3)*rho,sqrt(1/3),sqrt(2/3)*rho,sqrt(2/3),rho,1),
nrow=6,ncol=6)

############################## Defining functions ############################
 if(stages==3){
  out <- 1 - pmvnorm(lower=c(-Inf,-Inf,-Inf,-Inf,-Inf,-Inf), upper=c(bound[1],
```

```
  bound[1],bound[2],bound[2],bound[3],bound[3]),
  mean=as.numeric(c(mean11,mean21,mean12,mean22,mean13,mean23)),sigma=sigma3)
 } ### end if stages=3
out
} ### end gs2.power

ngs2.power <- function(n,stages,rho,diff,bound,power,s1,s2){
 gs2.power(n=n,stages=stages,rho=rho,diff=diff,bound=bound,s1=s1,s2=s2)[1]-power
} ### end ng2.power
```

Appendix E: GSD simulation program

```
##########################################################################
##This program computes the boundaries at each stage and the maximum sample
##sample size. It also performs the test of hypothesis and append the results
##at the end.
##########################################################################
## Load the packages below if needed
    library(lattice)
    library(mvtnorm)

## The program uses the source file "call" which contains the mean vector,
## the variance covariance matrix and the multivariate normal probability
## function. The progran also uses the source file "coded" which contains
## the main program to compute the boundaries

    source("./coded.R") # see appendix A
    source("./call.R") # see Appendix B
```

```
##### Before stage 1 the following parameters are set up ####################

stages <- 3        # number of stages
rho <- 0.5         # correlation
diff <- 0.5        # treatement effect
alpha <- 0.0125    #  nominal alpha
s1 <- 1.5          # sigma 1
s2 <- 1     # sigma 2
cv <- qnorm(1-alpha/2)  # calculate crical value
power <- 0.8 # target power
t10 <- 1/3  # initial time at stage 1
t20 <- 2/3  # initial time at stage 2
t30 <- 3/3 # initial time at stage 3

######### calculate boundaries before stage 1#################################
## calculate boundary at stage 1
spendfunct1 <- 2 - 2*pnorm(cv/sqrt(t10))
B10 <- qnorm(1-spendfunct1)

## calculate boundary at stage 2 using program  "coded"
# Input
#############################################################
tcur<- t20 # Current value of t
tprev<-c(t10) # Previous values of t go in
```

```
# parentheses.
cprev<-c(B10) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t20)) # Cumulative alpha up to current look,
# alpha_*(tcur)
############################################################
other<-c(tcur,tprev,cprev,cumulal)
B20 <- findroot(rename,-7,7,other)

## calculate boundary at stage 3 using program  "coded"
# Input
############################################################
tcur<-t30 # Current value of t
tprev<-c(t10,t20) # Previous values of t go in
# parentheses.
cprev<-c(B10,B20) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t30)) # Cumulative alpha up to current look,
# alpha_*(tcur)
############################################################
other<-c(tcur,tprev,cprev,cumulal)
B30 <- findroot(rename,-7,7,other) # boundary

boundary0 <- c(B10,B20,B30) # initial boundaries
```

```
## Calculate the initial maximum sample size
n0 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,stages=stages,rho=rho,
diff=diff,bound=boundary0,power=power,s1=s1,s2=s2)$root,0),silent=TRUE)
nmax0 <- n0*3

############ Simulate data for stage 1 ############################
################# simulation set-up ################################
nsim <- 100000 ### number of simulation
set.seed(1) ### number of seeds
results <- matrix(0,nrow=nsim,ncol=8) #### append file set up

###set up a file where number of rejection and expected sample size at each
stage are stored
colnames(results) <- c("efficacy1","expsample1","efficacy2","expsample2",
"efficacy3","expsample3","efficacy5","expsample5")

#### set up a file where  sigmas and maximum sample size are stored
results1 <- matrix(0,nrow=nsim,ncol=21)
colnames(results1) <- c("i","rho","s1","s2","nmax0","nrho1","ns11","ns21",
"nmax1","nrho2","ns12","ns22","nmax2","nrho3","ns13","ns23","nmax3","nrho4"
,"ns14","ns24","nmax4")

############################## Run simulations ####################
```

```
for (i in 1:nsim){
results1[i,1] <- i
results1[i,2] <- rho
results1[i,3] <- s1
results1[i,4] <- s2
results1[i,5] <- nmax0

#### initiate ####
istop <- 0
if (istop==0){

############################## stage 1 ###################
trurho <- 0.5 # true correlation
trus1 <- 1.5 # true sigma 1
trus2 <- 1.5 # true sigma 2
nsigma  <- matrix(c(trus1^2,trus1*trus2*trurho,trus1*trus2*trurho,
trus2^2),nrow=2,ncol=2) # true variance-covariance matrix
Tsample1 <- rmvnorm(nmax0/3,c(0.5,0.5),nsigma) # simulated data for E group
Csample1 <- rmvnorm(nmax0/3,c(0,0),nsigma) # simulated data for C group

## Estimate nuisance parameters
nrho1 <- cor(c(Tsample1[,1],Csample1[,1]),c(Tsample1[,2],Csample1[,2]))
# estimate rho
ns11 <- sqrt(var(c(Tsample1[,1],Csample1[,1]))) # estimate sigma 1
```

```
ns21 <- sqrt(var(c(Tsample1[,2],Csample1[,2]))) # estimate sigma 2
results1[i,6] <- nrho1 # append results
results1[i,7] <- ns11 # append results
results1[i,8] <- ns21 # append results

## Estimate sample size at stage 1
n1 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,stages=stages,rho=nrho1,
diff=diff,bound=boundary0,power=power,s1=ns11,s2=ns21)$root,0),silent=TRUE)
nmax1 <- n1*3 # maximum sample size
results1[i,9] <- nmax1 # append results

## calculate information fraction and type I error spent at stage 1
t11 <- nmax0/(3*nmax1) # information fraction at stage 1
spendfunct11 <- 2 - 2*pnorm(cv/sqrt(t11)) # type I error at stage 1
B11 <- qnorm(1-spendfunct11) # boundary at stage 1

## perform test of hypotheses
z11 <- (mean(Tsample1[,1])-mean(Csample1[,1]))/(ns11*(sqrt(2/(nmax0/3))))
    # z test  endpoint 1
z21 <- (mean(Tsample1[,2])-mean(Csample1[,2]))/(ns21*(sqrt(2/(nmax0/3))))
    # z test endpoint 2
efficacy1 <- as.numeric(z11>=(B11)|z21>=(B11)) # compare Z and boundary B11
results[i,1] <- as.numeric(z11>=(B11)|z21>=(B11))
```

```
if(efficacy1==1){
results[i,2] <- nmax0/3
istop <- 1
} #accept or reject
else {

################################## stage 2 ################################
# simulate more data
Tsample22 <- rmvnorm(((2*nmax1)/3)-(nmax0/3),c(0.5,0.5),nsigma)
Csample22 <- rmvnorm(((2*nmax1)/3)-(nmax0/3),c(0,0),nsigma)
Tsample2 <- rbind(Tsample1,Tsample22)
Csample2 <- rbind(Csample1,Csample22)

#Estimate nuisance parameters
nrho2 <- cor(c(Tsample2[,1],Csample2[,1]),c(Tsample2[,2],Csample2[,2]))
ns12 <- sqrt(var(c(Tsample2[,1],Csample2[,1])))
ns22 <- sqrt(var(c(Tsample2[,2],Csample2[,2])))

#caculate z statistics
z12 <- (mean(Tsample2[,1])-mean(Csample2[,1]))/(ns12*(sqrt(2/((2*nmax1)/3))))
z22 <- (mean(Tsample2[,2])-mean(Csample2[,2]))/(ns22*(sqrt(2/((2*nmax1)/3))))
results1[i,10] <- nrho2
results1[i,11] <- ns12
results1[i,12] <- ns22
```

```
####################Adjusting Old boundary ##############################
t12 <- nmax0/(3*nmax1)
cond1a <- as.numeric((t12 < 0.091))
if (cond1a==1){t12 <- 0.091}
spendfunct12 <- 2 - 2*pnorm(cv/sqrt(t12))
B12 <- qnorm(1-spendfunct12)
t22 <- (2*nmax1)/(3*nmax1) # information fraction at stage 2
cont <- as.numeric(t22>=1)
if (cont==1) {
t22 <- 1
istop <- 1
}
else {

# Input
#################################################################
tcur<-t22 # Current value of t
tprev<-c(t12) # Previous values of t go in
# parentheses.
cprev<-c(B12) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t22)) # Cumulative alpha up to current look,
# alpha_*(tcur)
```

```
#############################################################
other<-c(tcur,tprev,cprev,cumulal)
B22 <- findroot(rename,-7,7,other) # boundary at stage 2

t32 <- nmax1/nmax1

# Input
#############################################################
tcur<-t32 # Current value of t
tprev<-c(t12,t22) # Previous values of t go in
# parentheses.
cprev<-c(B12,B22) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t32)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#############################################################
other<-c(tcur,tprev,cprev,cumulal)
B32 <- findroot(rename,-7,7,other)

boundary1 <- c(B12,B22,B32) # adjusted boundaries

##### calculate sample size at stage 2 ########
n2 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,stages=stages,rho=nrho2,
diff=diff,bound=boundary1,power=power,s1=ns12,s2=ns22)$root,0),silent=TRUE)
```

```
nmax2 <- n2*3 # max sample size
results1[i,13] <- nmax2 # append results
efficacy2 <- as.numeric(z12>=(B22)|z22>=(B22)) # compare z and boundar B22
results[i,3] <- as.numeric(z12>=(B22)|z22>=(B22)) # append

if(efficacy2==1){
results[i,4] <- (2*nmax1)/3
istop <- 1
} # accept or reject
else {

##############################  stage 3 ################################
######Adusting of old boundaries##########

t13 <- nmax0/(3*nmax2)
spendfunct13 <- 2 - 2*pnorm(cv/sqrt(t13)) # type I erro spent
B13 <- qnorm(1-spendfunct13)
t23 <- (2*nmax1)/(3*nmax2)
t23 <- (2*nmax1)/(3*nmax2)

# Input
#################################################################
tcur<-t23 # Current value of t
tprev<-c(t13) # Previous values of t go in
```

```
# parentheses.
cprev<-c(B13) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t23)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#############################################################
other<-c(tcur,tprev,cprev,cumulal)
B23 <- findroot(rename,-7,7,other)

#print(B23)
t33 <- nmax2/nmax2

# Input
#############################################################
tcur<-t33 # Current value of t
tprev<-c(t13,t23) # Previous values of t go in
# parentheses.
cprev<-c(B13,B23) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t33)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#############################################################
other<-c(tcur,tprev,cprev,cumulal)
B33 <- findroot(rename,-7,7,other)
```

```
boundary2 <- c(B13,B23,B33) # adjusted boundaries at stage 3


# simulate data sta stage 3
Tsample33 <- rmvnorm(nmax2-(2*nmax1)/3,c(0.5,0.5),nsigma)
Csample33 <- rmvnorm(nmax2-(2*nmax1)/3,c(0,0),nsigma)
Tsample3 <- rbind(Tsample2,Tsample33)
Csample3 <- rbind(Csample2,Csample33)


#Estimate nuisance parameters
nrho3 <- cor(c(Tsample3[,1],Csample3[,1]),c(Tsample3[,2],Csample3[,2]))
    # estimate rho
ns13 <- sqrt(var(c(Tsample3[,1],Csample3[,1]))) # estimate sigma
ns23 <- sqrt(var(c(Tsample3[,2],Csample3[,2]))) # estimate sigma
# z statistics
z13 <- (mean(Tsample3[,1])-mean(Csample3[,1]))/(ns13*(sqrt(2/(nmax2))))
     # z statistic for endpoint 1
z23 <- (mean(Tsample3[,2])-mean(Csample3[,2]))/(ns23*(sqrt(2/(nmax2))))
    # z statistic for endpoint 2
# append estimate results
results1[i,14] <- nrho3
results1[i,15] <- ns13
results1[i,16] <- ns23
n3 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,stages=stages,rho=nrho3,
```

```
diff=diff,bound=boundary2,power=power,s1=ns13,s2=ns23)$root,0),silent=TRUE)
nmax3 <- n3*3 # max sample size
results1[i,17] <- nmax3

################ Adjust old Boundary ###############
t14 <- nmax0/(3*nmax3)
cond3 <- as.numeric(((t20 -t14) < 0.091) | (t14 < 0.091))
if (cond3==1) {t14 <- 0.091}
spendfunct14 <- 2 - 2*pnorm(cv/sqrt(t14))
B14 <- qnorm(1-spendfunct14)
t24 <- (2*nmax1)/(3*nmax3)
conti <- as.numeric(t24>=1)
if (conti==1) {
t24 <- 1
istop <- 1
}
else {

# Input
#################################################################
tcur<-t24 # Current value of t
tprev<-c(t14) # Previous values of t go in
# parentheses.
cprev<-c(B14) # Previous boundary values go in
```

```
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t24)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B24 <- findroot(rename,-7,7,other)

t34 <- (3*nmax2)/(3*nmax3)

# Input
#################################################################
tcur<-t34 # Current value of t
tprev<-c(t14,t24) # Previous values of t go in
# parentheses.
cprev<-c(B14,B24) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t34)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B34 <- findroot(rename,-7,7,other)

boundary3 <- c(B14,B24,B34) # new boundaries to use at stage 4
```

```
efficacy3 <- as.numeric(z13>=(B34) | z23>=(B34))
results[i,5] <- as.numeric(z13>=(B34) | z23>=(B34))

if (efficacy3==1){
results[i,6] <- nmax2
istop <- 1
}
else {

if (t34==1) {
istop <- 1
}

else {

#######################################stage 4#################################
#simulate data
Tsample44 <- rmvnorm((nmax3-nmax2),c(0.5,0.5),nsigma)
Csample44 <- rmvnorm((nmax3-nmax2),c(0,0),nsigma)
Tsample4 <- rbind(Tsample3,Tsample44)
Csample4 <- rbind(Csample3,Csample44)
#Estimate nuisance parameters
nrho4 <- cor(c(Tsample4[,1],Csample4[,1]),c(Tsample4[,2],Csample4[,2]))
     # estimate rho
```

```
ns14 <- sqrt(var(c(Tsample4[,1],Csample4[,1]))) # estimate sigma 1
ns24 <- sqrt(var(c(Tsample4[,2],Csample4[,2]))) # estimate sigma 2
results1[i,18] <- nrho4
results1[i,19] <- ns14
results1[i,20] <- ns24
# Sample size at stage 4
n4 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,stages=stages,rho=nrho4,
diff=diff,bound=boundary3,power=power,s1=ns14,s2=ns24)$root,0),silent=TRUE)
nmax4 <- n4*3 # max sample size
results1[i,21] <- nmax4

t4 <- 1 # information time

# Input
#################################################################
tcur<-t4 # Current value of t
tprev<-c(t14,t24,t34) # Previous values of t go in
# parentheses.
cprev<-c(B14,B24,B34) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t4)) # Cumulative alpha up to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
```

```
B4 <- findroot(rename,-7,7,other) # boundary at stage 4

z14 <- (mean(Tsample4[,1])-mean(Csample4[,1]))/(ns14*(sqrt(2/(nmax4))))
z24 <- (mean(Tsample4[,2])-mean(Csample4[,2]))/(ns24*(sqrt(2/(nmax4))))
efficacy4 <- as.numeric(z14>=(B4) | z24>=(B4))
results[i,7] <- as.numeric(z14>=(B4) | z24>=(B4))

if(efficacy4==1){
results[i,8] <- nmax4
istop <- 1}

} # end sim
} # end istop
} # end of look 1
} # end t22
} # end look 2
} # end of t23
} # end of t24
} # end of stage 3

##################  append results and results1 ##############
write.table(results,file="./guess05true05alt.txt",sep="\t",eol="\n",
col.names=TRUE,na = "NA",row.names=FALSE)
```

```
write.table(results1,file="./guess05true05alt1.txt",sep="\t",eol="\n",
col.names=TRUE,na = "NA",row.names=FALSE)
```

Appendix F: GSD inverse normal combination test simulation program

```
library(lattice)
#library(ldbounds)
library(mvtnorm)

source("./call.R")
source("./coded.R")
#setwd("H:/IPS/Results")

stages <- 3
rho <- 0.5
diff <- 0.5
alpha <- 0.0125
s1 <- 1.5
s2 <- 1
cv <- qnorm(1-alpha/2)
```

```
power <- 0.8
t10 <- 1/3
t20 <- 2/3
t30 <- 3/3

## calculate boundary at look 1###

spendfunct1 <- 2 - 2*pnorm(cv/sqrt(t10))

B10 <- qnorm(1-spendfunct1)

## calculate boundary at look 2##

# Input
################################################################
tcur<- t20 # Current value of t
tprev<-c(t10) # Previous values of t go in
# parentheses.
cprev<-c(B10) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t20)) # Cumulative alpha up to
current look,
# alpha_*(tcur)
################################################################
```

```
other<-c(tcur,tprev,cprev,cumulal)
B20 <- findroot(rename,-7,7,other)


## calculate boundary at look 3##

# Input
################################################################
tcur<-t30 # Current value of t
tprev<-c(t10,t20) # Previous values of t go in
# parentheses.
cprev<-c(B10,B20) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t30)) # Cumulative alpha up to
current look,
# alpha_*(tcur)
################################################################
other<-c(tcur,tprev,cprev,cumulal)
B30 <- findroot(rename,-7,7,other)

boundary0 <- c(B10,B20,B30)
#print(boundary2)

## Sample size
```

```
#n0 <- round(uniroot(gs2.power,lower=1,upper=100000,rho=rho,
diff=diff,bound=boundary0,stages,s1,s2)$root)
n0 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,
stages=stages,
rho=rho,diff=diff,bound=boundary0,power=power,s1=s1,
s2=s2)$root,0),silent=TRUE)
nmax0 <- n0*3



# simulation set-up
nsim <- 100000
set.seed(1)
results <- matrix(0,nrow=nsim,ncol=8)
colnames(results) <- c("efficacy1","expsample1",
"efficacy2","expsample2",
"efficacy3","expsample3","efficacy5","expsample5")
results1 <- matrix(0,nrow=nsim,ncol=21)
colnames(results1) <- c("i","rho","s1","s2","nmax0",
"nrho1","ns11","ns21",
"nmax1","nrho2","ns12","ns22","nmax2","nrho3","ns13",
"ns23","nmax3","nrho4","ns14","ns24","nmax4")

# run simulations
```

```
for (i in 1:nsim){

results1[i,1] <- i
results1[i,2] <- rho
results1[i,3] <- s1
results1[i,4] <- s2
results1[i,5] <- nmax0



# initiate
istop <- 0

if (istop==0){

############################## Look1 ###################

trurho <- 0
trus1 <- 1
trus2 <- 1.5
nsigma  <- matrix(c(trus1^2,trus1*trus2*trurho,
trus1*trus2*trurho,trus2^2),nrow=2,ncol=2)

Tsample1 <- rmvnorm(nmax0/3,c(0,0),nsigma)
```

```
Csample1 <- rmvnorm(nmax0/3,c(0,0),nsigma)

##Blinded Sample Size##
nrho1 <- cor(c(Tsample1[,1],Csample1[,1]),
c(Tsample1[,2],Csample1[,2]))
ns11 <- sqrt(var(c(Tsample1[,1],Csample1[,1])))
ns21 <- sqrt(var(c(Tsample1[,2],Csample1[,2])))



condit11 <- as.numeric(nrho1>1)
if (condit11==1) { nrho1 <- 1}

results1[i,6] <- nrho1
results1[i,7] <- ns11
results1[i,8] <- ns21



n1 <- try(round(uniroot(ngs2.power,lower=1,
upper=10000,
stages=stages,rho=nrho1,diff=diff,bound=boundary0,
power=power,s1=ns11,s2=ns21)$root,0),silent=TRUE)
nmax1 <- n1*3
dfe1 <- (2*(nmax0/3))-2
results1[i,9] <- nmax1
```

```
t11 <- nmax0/(3*nmax1)
cond1 <- as.numeric(((t20 - t11) < 0.091)|
(t11 < 0.091))
if (cond1==1){t11 <- 0.091}
spendfunct11 <- 2 - 2*pnorm(cv/sqrt(t11))
B11 <- qnorm(1-spendfunct11)


Tsample11 <- rmvnorm(nmax1/3,c(0,0),nsigma)
Csample11 <- rmvnorm(nmax1/3,c(0,0),nsigma)

z11 <- (mean(Tsample11[,1])-mean(Csample11[,1]))
/(ns11*(sqrt(2/(nmax1/3))))
z21 <- (mean(Tsample11[,2])-mean(Csample11[,2]))
/(ns21*(sqrt(2/(nmax1/3))))

dfe11 <- (2*(nmax1/3))-2

p1 <- (1-pt(z11,dfe11))
p2 <- (1-pt(z21,dfe11))

zz11= qnorm(1-p1)
zz21= qnorm(1-p2)
```

```
efficacy1 <- as.numeric(zz11>=(B11)|zz21>=(B11))
#futility1 <- as.numeric(zz11<(-B11) & zz21<(-B11))
results[i,1] <- as.numeric(zz11>=(B11)|zz21>=(B11))

if(efficacy1==1){
results[i,2] <- nmax0/3
istop <- 1
}

else {

################################## Look 2 ####################


ttt <- ((2*nmax1)/3)-(nmax0/3)
kat <- as.numeric(ttt<5)
if (kat==1) {ttt <-7}

Tsample2 <- rmvnorm(ttt,c(0,0),nsigma)
Csample2 <- rmvnorm(ttt,c(0,0),nsigma)

#Tsample2 <- rmvnorm(((2*nmax1)/3)-(nmax0/3),c(0,0),nsigma)
#Csample2 <- rmvnorm(((2*nmax1)/3)-(nmax0/3),c(0,0),nsigma)
```

```
#J <- ((2*nmax1)/3)-(nmax0/3)



#Tsample2 <- rbind(Tsample1,Tsample22)
#Csample2 <- rbind(Csample1,Csample22)

#########################Blinded Sample Size##
nrho2 <- cor(c(Tsample2[,1],Csample2[,1]),c(Tsample2[,2],
Csample2[,2]))
ns12 <- sqrt(var(c(Tsample2[,1],Csample2[,1])))
ns22 <- sqrt(var(c(Tsample2[,2],Csample2[,2])))

condit22 <- as.numeric(nrho2>1)
if (condit22==1) { nrho2 <- 1}

#condit222 <- as.numeric(nrho2<0)
#if (condit222==1) { nrho2 <- 0}

#z12 <- (mean(Tsample2[,1])-mean(Csample2[,1]))/(ns12*
(sqrt(2/((2*nmax1-nmax0)/3))))
#z22 <- (mean(Tsample2[,2])-mean(Csample2[,2]))/(ns22*
(sqrt(2/((2*nmax1-nmax0)/3))))
```

```
z12 <- (mean(Tsample2[,1])-mean(Csample2[,1]))/(ns12*
(sqrt(2/((ttt)))))
z22 <- (mean(Tsample2[,2])-mean(Csample2[,2]))/(ns22*
(sqrt(2/((ttt)))))

#### calculate degree of freedom, p-values and IVN tests ####

#dfe2 = 2*((2*nmax1-nmax0)/3)-2
#dfe2 = 2*ttt-2

#p11 <- (1-pnorm(z12))
#p22 <- (1-pnorm(z22))

#zz12=(1/sqrt(2))*(qnorm(1-p1)+qnorm(1-p11))
#zz22=(1/sqrt(2))*(qnorm(1-p2)+qnorm(1-p22))



results1[i,10] <- nrho2
results1[i,11] <- ns12
results1[i,12] <- ns22

####################Adjusting Old boundary ###############
```

```
t12 <- nmax0/(3*nmax1)
cond1a <- as.numeric((t12 < 0.091))
if (cond1a==1){t12 <- 0.091}
spendfunct12 <- 2 - 2*pnorm(cv/sqrt(t12))
B12 <- qnorm(1-spendfunct12)

t22 <- (2*nmax1)/(3*nmax1)
cont <- as.numeric(t22>=1|t22<=t12|(t22 - t12) < 0.091)
if (cont==1) {
#t22 <- 1
istop <- 1
}

else {

# Input
################################################################
tcur<-t22 # Current value of t
tprev<-c(t12) # Previous values of t go in
# parentheses.
cprev<-c(B12) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t22)) # Cumulative alpha up
to current look,
```

```
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B22 <- findroot(rename,-7,7,other)

t32 <- nmax1/nmax1


# Input
#################################################################
tcur<-t32 # Current value of t
tprev<-c(t12,t22) # Previous values of t go in
# parentheses.
cprev<-c(B12,B22) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t32)) # Cumulative alpha up
 to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B32 <- findroot(rename,-7,7,other)


boundary1 <- c(B12,B22,B32)
```

```
n2 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,
stages=stages,rho=nrho2,diff=diff,bound=boundary1,
power=power,s1=ns12,s2=ns22)$root,0),silent=TRUE)

nmax2 <- n2*3



tttt <- ((2*nmax2)/3)-(nmax1/3)
katt <- as.numeric(tttt<5)
if (katt==1) {tttt <-7}

Tsample22 <- rmvnorm(tttt,c(0,0),nsigma)
Csample22 <- rmvnorm(tttt,c(0,0),nsigma)



z12 <- (mean(Tsample22[,1])-mean(Csample22[,1]))
/(ns12*(sqrt(2/((tttt)))))
z22 <- (mean(Tsample22[,2])-mean(Csample22[,2]))
/(ns22*(sqrt(2/((tttt)))))

dfe22 = 2*tttt-2
```

```
p11 <- (1-pt(z12,dfe22))
p22 <- (1-pt(z22,dfe22))


zz12=(1/sqrt(2))*(qnorm(1-p1)+qnorm(1-p11))
zz22=(1/sqrt(2))*(qnorm(1-p2)+qnorm(1-p22))


kkk <- nmax2- ((2*nmax1)/3)
condaaa <- as.numeric(kkk <= 5)
if (condaaa==1) {kkk <- 10}

#else{

results1[i,13] <- nmax2
efficacy2 <- as.numeric(zz12>=(B22)|zz22>=(B22))
#futility2 <- as.numeric(z12<(-B22) & z22<(-B22))
results[i,3] <- as.numeric(zz12>=(B22)|zz22>=(B22))


if(efficacy2==1){
results[i,4] <- (2*nmax1)/3
istop <- 1
}
```

```
else {

############################## Look 3 ###############

######Adusting of old boundaries##########

t13 <- nmax0/(3*nmax2)
cond2 <- as.numeric(((t20 - t13) < 0.091) |
(t13 <0.091))
if  (cond2==1) {t13 <- 0.091}
spendfunct13 <- 2 - 2*pnorm(cv/sqrt(t13))
B13 <- qnorm(1-spendfunct13)


t23 <- (2*nmax1)/(3*nmax2)
conda <- as.numeric(t23>=1 | t23<=t13|(t23 - t13)
< 0.091)
if (conda==1) {istop <- 1}

else {

#condabis <- as.numeric(((t23 - t13) < 0.091) |
(t23 <0.091))
```

```
#if (condabis==1) {t23 <- 0.091}



# Input
############################################################
tcur<-t23 # Current value of t
tprev<-c(t13) # Previous values of t go in
# parentheses.
cprev<-c(B13) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t23)) # Cumulative
alpha up to current look,
# alpha_*(tcur)
############################################################
other<-c(tcur,tprev,cprev,cumulal)
B23 <- findroot(rename,-7,7,other)

#print(B23)
t33 <- nmax2/nmax2



# Input
############################################################
```

```
tcur<-t33 # Current value of t
tprev<-c(t13,t23) # Previous values of t go in
# parentheses.
cprev<-c(B13,B23) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t33)) # Cumulative
alpha up to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B33 <- findroot(rename,-7,7,other)

boundary2 <- c(B13,B23,B33)

Tsample3 <- rmvnorm(kkk,c(0,0),nsigma)
Csample3 <- rmvnorm(kkk,c(0,0),nsigma)
#Tsample3 <- rmvnorm(nmax2-(2*nmax1)/3,c(0,0),nsigma)
#Csample3 <- rmvnorm(nmax2-(2*nmax1)/3,c(0,0),nsigma)
#Tsample3 <- rbind(Tsample2,Tsample33)
#Csample3 <- rbind(Csample2,Csample33)

##Blinded Sample Size##
nrho3 <- cor(c(Tsample3[,1],Csample3[,1]),
c(Tsample3[,2],Csample3[,2]))
```

```
ns13 <- sqrt(var(c(Tsample3[,1],Csample3[,1])))
ns23 <- sqrt(var(c(Tsample3[,2],Csample3[,2])))

condit33 <- as.numeric(nrho3>1)
if (condit33==1) {nrho3 <- 1}



#condit333 <- as.numeric(nrho3<0)
#if (condit333==1) { nrho3 <- 0}




#z13 <- (mean(Tsample3[,1])-mean(Csample3[,1]))
/(ns13*(sqrt(2/(kkk))))
#z23 <- (mean(Tsample3[,2])-mean(Csample3[,2]))
/(ns23*(sqrt(2/(kkk))))

#z13 <- (mean(Tsample3[,1])-mean(Csample3[,1]))
/(ns13*(sqrt(2/(nmax2-((2*nmax1)/3)))))
#z23 <- (mean(Tsample3[,2])-mean(Csample3[,2]))
/(ns23*(sqrt(2/(nmax2-((2*nmax1)/3)))))


#dfe3 = 2*(nmax2-(2*nmax1)/3)-2
```

```
#dfe3 = 2*kkk-2


#p13 <- (1-pt(z13,dfe3))
#p23 <- (1-pt(z23,dfe3))


#zz13=(1/sqrt(3))*(qnorm(1-p1)+qnorm(1-p11)+qnorm(1-p13))
#zz23=(1/sqrt(3))*(qnorm(1-p2)+qnorm(1-p22)+qnorm(1-p23))


######
results1[i,14] <- nrho3
results1[i,15] <- ns13
results1[i,16] <- ns23


n3 <- try(round(uniroot(ngs2.power,lower=1,upper=10000,
stages=stages,rho=nrho3,diff=diff,bound=boundary2,
power=power,s1=ns13,s2=ns23)$root,0),silent=TRUE)
nmax3 <- n3*3
results1[i,17] <- nmax3



kkkk <- nmax3- ((2*nmax2)/3)
condaaaa <- as.numeric(kkkk <= 5)
if (condaaaa==1) {kkkk <- 10}
```

```
Tsample33 <- rmvnorm(kkkk,c(0,0),nsigma)
Csample33 <- rmvnorm(kkkk,c(0,0),nsigma)


z13 <- (mean(Tsample33[,1])-mean(Csample33[,1]))
/(ns13*(sqrt(2/(kkkk))))
z23 <- (mean(Tsample33[,2])-mean(Csample33[,2]))
/(ns23*(sqrt(2/(kkkk))))

#z13 <- (mean(Tsample3[,1])-mean(Csample3[,1]))
/(ns13*(sqrt(2/(nmax2-((2*nmax1)/3)))))
#z23 <- (mean(Tsample3[,2])-mean(Csample3[,2]))
/(ns23*(sqrt(2/(nmax2-((2*nmax1)/3)))))


#dfe3 = 2*(nmax2-(2*nmax1)/3)-2
dfe33 = 2*kkkk-2

p13 <- (1-pt(z13,dfe33))
p23 <- (1-pt(z23,dfe33))

zz13=(1/sqrt(3))*(qnorm(1-p1)+qnorm(1-p11)
+qnorm(1-p13))
```

```
zz23=(1/sqrt(3))*(qnorm(1-p2)+qnorm(1-p22)
+qnorm(1-p23))
################New Boundary###############


t14 <- nmax0/(3*nmax3)
cond3 <- as.numeric(((t20 -t14) < 0.01)
| (t14 < 0.091))
if (cond3==1) {t14 <- 0.091}
spendfunct14 <- 2 - 2*pnorm(cv/sqrt(t14))
B14 <- qnorm(1-spendfunct14)


t24 <- (2*nmax1)/(3*nmax3)
conti <- as.numeric(((t24 -t14) < 0.091)
| (t24>=1) | t24 <=t14)
if (conti==1) {
#t24 <- 1
istop <- 1
}

else {

# Input
```

```
##########################################################
tcur<-t24 # Current value of t
tprev<-c(t14) # Previous values of t go in
# parentheses.
cprev<-c(B14) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t24)) # Cumulative
 alpha up to current look,
# alpha_*(tcur)
##########################################################
other<-c(tcur,tprev,cprev,cumulal)
B24 <- findroot(rename,-7,7,other)

t34 <- (3*nmax2)/(3*nmax3)
contii <- as.numeric(t34>=1 |((t34 - t24) < 0.10))
#contii <- as.numeric(t34>=1)
if (contii==1) {
t34 <- 1
}

# Input
##################################################################
tcur<-t34 # Current value of t
tprev<-c(t14,t24) # Previous values of t go in
```

```
# parentheses.
cprev<-c(B14,B24) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t34)) # Cumulative
alpha up to current look,
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)
B34 <- findroot(rename,-7,7,other)

boundary3 <- c(B14,B24,B34)



efficacy3 <- as.numeric(zz13>=(B34) | zz23>=(B34))
#futility3 <- as.numeric(z13<(-B34) & z23<(-B34))
results[i,5] <- as.numeric(zz13>=(B34) | zz23>=(B34))

if (efficacy3==1){
results[i,6] <- nmax2
istop <- 1
}

else {
```

```
if (t34==1) {
istop <- 1

}

else {
#####################Look 4#################################

ppp <- nmax3-nmax2

condit444 <- as.numeric(ppp < 5)

if (condit444==1) {ppp <- 7}

Tsample4 <- rmvnorm(ppp,c(0,0),nsigma)
Csample4 <- rmvnorm(ppp,c(0,0),nsigma)

#Tsample4 <- rmvnorm((nmax3-nmax2),c(0,0),nsigma)
#Csample4 <- rmvnorm((nmax3-nmax2),c(0,0),nsigma)
#Tsample4 <- rbind(Tsample3,Tsample44)
#Csample4 <- rbind(Csample3,Csample44)
#Tsample4 <- rmvnorm(nmax3,c(0.5,0.5),nsigma)
#Csample4 <- rmvnorm(nmax3,mu0,nsigma)
```

```
#Blided sample size

nrho4 <- cor(c(Tsample4[,1],Csample4[,1]),c
(Tsample4[,2],Csample4[,2]))
ns14 <- sqrt(var(c(Tsample4[,1],Csample4[,1])))
ns24 <- sqrt(var(c(Tsample4[,2],Csample4[,2])))


#condit444 <- as.numeric(nrho4<0)

#if (condit444==1) {nrho4 <- 0}

condit44 <- as.numeric(nrho4>1)

if (condit44==1) { nrho4 <- 1}
results1[i,18] <- nrho4
results1[i,19] <- ns14
results1[i,20] <- ns24

#cojeka <- as.numeric((nmax3-nmax2)<=3)
# if (cojeka==1) {istop <- 1}
#else{

n4 <- try(round(uniroot(ngs2.power,lower=1,upper=10000
```

```
,stages=stages,rho=nrho4,diff=diff,bound=boundary3
,power=power,s1=ns14,s2=ns24)$root,0),silent=TRUE)
nmax4 <- n4*3
results1[i,21] <- nmax4


#cond333 <- as.numeric(nmax3<=nmax2)
# if (cond333==1) {istop <- 1}

#else {

###
t4 <- 1
#print(t4)

# Input
################################################################
tcur<-t4 # Current value of t
tprev<-c(t14,t24,t34) # Previous values of t go in
# parentheses.
cprev<-c(B14,B24,B34) # Previous boundary values go in
# parentheses.
cumulal<- 2 - 2*pnorm(cv/sqrt(t4)) # Cumulative
alpha up to current look,
```

```
# alpha_*(tcur)
#################################################################
other<-c(tcur,tprev,cprev,cumulal)


B4 <- findroot(rename,-7,7,other)

sss <- nmax4-nmax3

condit4444 <- as.numeric(sss <= 3)

if (condit4444==1) {istop <- 1}

else {

Tsample44 <- rmvnorm(sss,c(0,0),nsigma)
Csample44 <- rmvnorm(sss,c(0,0),nsigma)


z14 <- (mean(Tsample44[,1])-mean(Csample44[,1]))
/(ns14*(sqrt(2/(sss))))
z24 <- (mean(Tsample44[,2])-mean(Csample44[,2]))
/(ns24*(sqrt(2/(sss))))
```

```
#z14 <- (mean(Tsample4[,1])-mean(Csample4[,1]))
/(ns14*(sqrt(2/(nmax3-nmax2))))
#z24 <- (mean(Tsample4[,2])-mean(Csample4[,2]))
/(ns24*(sqrt(2/(nmax3-nmax2))))

########calculate p-value,

#dfe4 = 2*(nmax3-nmax2)-2
dfe4 = 2*sss-2

p14 <- (1-pt(z14,dfe4))
p24 <- (1-pt(z24,dfe4))

zz14= (1/sqrt(4))*(qnorm(1-p1)+qnorm(1-p11)
+qnorm(1-p13)+qnorm(1-p14))
zz24= (1/sqrt(4))*(qnorm(1-p2)+qnorm(1-p22)
+qnorm(1-p23)+qnorm(1-p24))

efficacy4 <- as.numeric(zz14>=(B4) | zz24>=(B4))
futility4 <- as.numeric(zz14<(-B4) & z24<(-B4))
results[i,7] <- as.numeric(zz14>=(B4) | zz24>=(B4))


if(efficacy4==1){
```

```
results[i,8] <- nmax4
istop <- 1}



} #end sim
} #end istop
} # end of look 1
} #end t22
} #end look 2
} #end of t23
} #end of t24
} #end of stage 3
}
 }
#}
#}
#}
write.table(results,file="./guess05true00null.txt"
,sep="\t",eol="\n",col.names=TRUE,na = "NA",row.names=FALSE)
write.table(results1,file="./guess05true00null1.txt"
,sep="\t",eol="\n",col.names=TRUE,na = "NA",row.names=FALSE)
```

Bibliography

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician* **46**, 261–283.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9–26.

Armitage, P., C. McPherson, and B. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society A* **132**, 235–244.

Bank, N., P. Bauer, and J. Röhmel (1996). On the power of fishers combination test for two stage sampling in the presence of nuisance parameters. *Biometrical* **38**, 25–37.

Basu, D. (1977). On the elimination of nuisance parameter. *American Statistical Association* **72**, 355–366.

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biom. und Inform. in Med* **20**, 130 – 148.

Bauer, P. (1989b). Sequential tests of hypotheses in consecutive trials. *Biometrical* **31**, 663 – 676.

Bauer, P. (1992). The choice of sequential boundaries based on the concept of power spending. *Biom. Und Inf. In Med* **20**, 130 – 148.

Bauer, P. and M. Kieser (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833 – 1848.

Bauer, P. and K. Kohne (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–41.

Bauer, P. and J. Röhmel (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bauer, P. and J. Röhmel (1995). An adaptive method for establishing a dose response relationship. *Statistics in Medicine* **14**, 1595 – 1607.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate:a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological) 57 (1)*, 289–300.

Birkett, M. A. and S. J. Day (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455 – 2463.

Bland, M. (2000). *An itroduction to Medical Statistics* (3 ed.). United Kingdom: Oxford University Press.

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management* (2 ed.). Thousand Oaks, CA: Sage Publications.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Ist. Super. di Sci. Econom. e Commerciali di Firenze* **8**, 1–62.

Borma, G. F., G. J. van der Wiltb, J. A. Kremerc, and G. A. Zielhuisa. A generalized concept of power helped to choose optimal endpoints in clinical trials.

Brannath, W., M. Posch, and P. Bauer (2002). Recursive combination tests. *American Statistical Association* **97**, 236 – 244.

Buchanan, R. W., D. G. Miriam Davis, M. F. Green, R. S. E. Keefe, A. C. Leon, K. H. Nuechterlein, T. Laughren, R. Levin, E. Stover, W. Fenton, and S. R. Marder (2005). A summary of the FDA-NIMH-MATRICS workshop on clinical trial design for neurocognitive drugs for schizophrenia. *Schizophrenia Bulletin* **31**, 1 – 19.

Castle, D., S. Wessely, G. Der, and R. Murray (1991). The incidence of operationally defined schizophrenia in camberwell, 1965-84. *The British Journal of Psychiatry* **159**, 790–794.

Chow, S.-C. and M. Chang (2006). *Adaptive design methods in clinical trials* (1 ed.). USA: Boca Raton: Chapman and Hall/CRC.

Chow, S. C., J. Shao, and H. Wang (2008). *Sample size calculations in clinical research* (**2** ed.). USA: Chapman & Hall/CRC Biostatistics series.

Coffey, C. S. and K. E. Muller (1999). Exact test size and power of a gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199 – 1214.

Coffey, C. S. and K. E. Muller (2001). Controlling test size while gaining the benefits of an internal pilot design. *Biometrics* **57**, 625–631.

Conaway, M. R. and G. Petroni (1995). Bivariate sequential designs for phase ii trials. *Biometrics* **51**, 565–664.

Cook, R. J. (1994). Interim monitoring of bivariate responses using repeated confidence intervals. *Controlled Clinical Trials* **15**, 187–200.

Cook, R. J. and V. T. Farewell (1994). Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics* **50**, 1146–1152.

Costa, L. M., J. Achten, R. N. Parsons, P. R. Edlin, P. Foguet, U. Prakash, R. D. Griffin, and Y. Adult (2012). Total hip arthroplasty versus resurfacing arthroplasty in the treatment

of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial. *British Medical Journal* **334**, 1–9.

Cui, L., H. M. J. Huang, and S.-J. Wang (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853 – 857.

Day, S. and D. Altman (2000). Blinding in clinical trials and other studies. *British Medical Journal* **321**, 19–26.

Degroot, M. H. and M. J. Schervish (2002). *Probability and Statistics* (**3** ed.). USA: Addison Wesley.

Demets, D. L., C. D. Furberg, and L. M. Friedman (2006). *Data Monitoring in Clinical Trials: A case Studies Approach* (1 ed.). USA: Springer.

Denne, J. S. and C. Jennison (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* **18**, 1575 – 1585.

Dmitrienko, A., G. Molenberghs, C. Chuang-Stein, and W. Offen (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide* (1 ed.). USA: SAS Institute Inc., Cary, NC, USA.

Ekenstierna, M. (2004). *Multiple comparison procedures based onmarginal p-values* (1 ed.). Uppsala University: U.U.D.M. Project Report.

Emerson, S. and T. Fleming (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905923.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551 – 1562.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (1 ed.). London: Oliver & Boyd.

Fleischhacker, W. W. and G. M. Goodwin (2009). Effectiveness as an outcome measure for treatment trials in psychiatry. *World Psychiatry* **8**, 23–27.

Friede, T. and M. Kieser (2001). A compararison of methods for adaptive sample size adjustment. *Statistics in Medicine* **20**, 3861 – 3873.

Friede, T. and M. Kieser (2002). On the inappropriateness of an em algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine* **21**, 165–176.

Friede, T. and M. Kieser (2003). Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine* **22**, 995–1007.

Friede, T. and M. Kieser (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal* **4**, 1–19.

Friede, T. and M. Kieser (2009). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics* **10**, 8–13.

Friede, T. and H. Schmidli (2010). Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statistics in Medecine* **29**, 1145–1156.

Geller, N. L. (2004). *Advances in clinical trial biostatistics* (1 ed.). New York: Marcel Dekker.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Computational and Graphical Statistics* **1**, 141–150.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, B. Bornkamp, and T. Hothorn (January 20, 2012). Multivariate normal and t distributions. *R-project*, 1–15.

Glimm, E., W. Maurer, and F. Bretz (2010). Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine* **29**, 219–228.

Gould, A. (1992). Interim analysis for monitoring clinical trials that do not materially affect the type i error rate. *Statistics in Medicine* **11**, 55–56.

Gould, A. and W. Shih (1992a). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Com Stat Theory Meth* **21**, 2833–53.

Gould, A. L. (2001). Sample size re-estimation: recent developements and practical considerations. *Statistics in Medecine* **20**, 2625–2643.

Gould, A. L. and W. J. Shih (1992b). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods* **21**, 2833 – 2853.

Hedges, L. V. and Olkin (1995). *Statistical Methods for Meta-Analysis* (1 ed.). New York: Academic Press.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800 – 802.

Hochberg, Y. and A. Tamhane (1987). *Multiple comparison procedures* (1 ed.). New York, USA: John Wiley and Sons, Inc.

Hochberg, Y. and A. Tamhane (2001). Multiple comparison methods. *Statistics in Medicine* **20**, 3861 – 3873.

Holm, S. (1979). Simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65 – 70.

Hommel, G. A stagewise rejective multiple test procedure based on a modified bonferonni test.

Hommel, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical* **25**, 423 – 430.

Hotelling, H. (1931). Procedures for comparing samples with multiple endpoints. *Annals of Mathematical Statistics* **2**, 360–378.

Huque, M. (2005). *Multiple Endpoint Testing in Clinical Trials Some Issues and Considerations*. Washington. DC: Industry/FDA Workshop.

Hwang, I. K., W. J. Shih, and J. S. De cani (1990). Group sequential designs using a family of type i probility spending functions. *Statistics in Medecine* **9**, 1439–1445.

ICH (2005). Ich harmonized tripartite guideline e9 (1999). statistical principles for clinical trials. *Statistics in Medicine* **18**, 1905– 1942.

Jennison, C. and B. Turnbull (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science* **5**, 299317.

Jennison, C. and B. Turnbull (1997). Group sequential analysis incorporating covariate information. *Americal Statistical Association* **92**, 1330–1341.

Jennison, C. and B. W. Turnbull (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety. *Biometrics* **49**, 741–752.

Jennison, C. and B. W. Turnbull (2000a). *Group Sequential Methods with application to clinical trials* (1 ed.). USA: Chapman and Hall/CRC.

Jennison, C. and B. W. Turnbull (2000b). *Group sequential methods with applications to clinical trial*. USA: Chapman & Hall/CRC.

Jennison, C. and B. W. Turnbull (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971 – 993.

Jennison, C. and B. W. Turnbull (2004). Adaptive re-design of clinical trials. *Paper presented at International Conference on Statistics in Health Sciences Nantes, France, June* $\mathbf{23 - 25}$.

Julious, S. A. (2004). Tutorial in biostatistics: Sample sizes for clinical trials with normal data. *Statist in Medecine* **23**, 1921–1986.

Kelly, P. J., M. R. Sooriyarachchi, N. Stallard, and S. Todd (2005). A practical comparison of group sequential and adaptive designs. *Biopharmaceutical Statistics* **15**, 719 – 738.

Kieser, M., P. Bauer, and W. Lehmacher (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical* **41**, 261 – 277.

Kieser, M. and T. Friede (2000). Re-calculating the sample size in internal pilot study designs with control of the type i error rate. *Statistics in Medicine* **19**, 901–911.

Kieser, M. and T. Friede (2003). Simple procedure for blinded sample size adjustment that do not affect the type i error rate. *Statistics in Medicine* **22**, 3571–3581.

Kim, K. and D. DeMets (1987). Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika* **74**, 149–154.

Kosorok, M. R., Y. Shi, and D. L. DeMets (2004). Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* **60**, 134–145.

Lan, K. and D. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**; **3**, 659–663.

Lan, K. and M. H. R Simon and (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics. Sequential Analysis* **1**, 207–219.

Legault, C. Analyzing multiple endpoints with a two-stages group sequential design in clinical trials.

Lehmacher, W. and G. Wassmer (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286–1290.

Leon, A. C. (2008). Implications of clinical trial design on sample size requirements. *Schizophrenia Bulletin* **34**, 664669.

Li, G., W. J. Shih, T. Xie, and J. Lu (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **3**, 277 – 287.

Liu, Q. and G. Y. H. Chi (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57**, 172 – 177.

Machin, D., S. Day, and S. Green (2004). *Texbook of clinical trial*. UK: Jon Wiley & Son, Ltd.

McPherson, K. and P. Armitage (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statitical Society* **SeriesA**, **134**, 15–25.

Mehta, C., P. Gao, D. Bhatt, and et al (2009). Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* **119**, 597605.

Mehta, C. and A. Tsiatis (2001). Flexible sample size consideration using information-based interim monitoring. *Drug Information* **35**, 1095112.

Miller, F. (2005). Variance estimation in clinical studies with interim sample size reestimation. *Biometrics* **61**, 355–361.

Mosteller, F. and R. Bush (1932). *Selected Quantitative Techniques* (1 ed.). Cambridge, Addison-Wesley: Handbook of Social Psychology.

Mosteller, F. and R. Bush (1954). *Selected quantitative techniques* (1 ed.). Cambridge: Handbook of social psychology.

Moye, L. A. (2003). *Multiple analyses in clinical trials: fundamentals for investigators* (1 ed.). New York: John Wiley & Sons.

Müller, H. H. and Schäfer (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886 – 891.

Muller, H.-H. and H. Schafer (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886–891.

Neuhäuser, M. (2006). How to deal with multiple endpoints in clinical trials. *Fundamental and Clinical Pharmacology* **20**, 515–523.

O'Brien, P. (1984). The generalization of student's ratio. *Biometrics* **40**, 1079–1087.

O'Brien, P. and T. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 459–456.

on Behalf of the MERIT-HF Study Group, T. I. S. C. (1997). Rational, design, and organisation of the metropol cr/xl randomized in intervention trial in heart failure (merit-hf). *American Journal of Cardiology* **80**, 54J–58J.

Pampallona, S. and A. Tsiatis (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Statistical Planning and Inference* **42**, 19–39.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *BMJ* **316**, 1236–1238.

Pocock, S., N. Geller, and A. Tsiatis (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **2**, 191199.

Pocock, S. J. (1982). Interim analysis for randomized clinical trials: The group sequential approach. *Biometrics* **38**, 153–162.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Elsevier Science* **18**, 530–545.

Pocock, S. J. (2004). *Clinical trial: a practical approach*. UK: Jon Wiley & Son, Ltd.

Pocock, S. J., M. D. Hughes, and R. J. Lee (1987). Statistical problems in the reporting of clinical trials. *New England Journal of Medicine* **317**, 426–432.

Posch, M. and P. Bauer (1999). Adaptive two stage designs and the conditional error function. *Biometrical* **41**, 689 – 696.

Proschan, M., Q. Liu, and S. Hunsberger (2003). Practical midcourse sample size modification in clinical trials. *Control Clincal Trials* **24**, 4–15.

Proschan, M. A. (2003). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Statist. Sinica* **13**, 163 – 177.

Proschan, M. A. (2005). An improved double sampling method based on the variance. *Biopharmaceutical Statistics* **15**, 559 – 574.

Proschan, M. A. (2009a). Sample size re-estimation in clinical trials. *Biometrical* **2**, 348 – 357.

Proschan, M. A. (2009b). Sample size re-estimation in clinical trials. *Biometrical Journal* **51**, 348–357.

Proschan, M. A. and S. A. Hunsberger (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315 – 1324.

Proschan, M. A., K. G. Lan, and J. T. Wittes (p.94,2006). *Statistical Monitoring of Clinical Trials: A Unified Approach* (1 ed.). USA: Springer.

Proschan, M. A. and J. Wittes (2000). An improved double sampling method based on the variance. *Biometrics* **56**, 1183 – 1187.

Quan, H., X. Luo, and T. Capizzi (2005). Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active treatment. *STATISTICS IN MEDICINE* **24**, 2151–2170.

Rom, D. A sequentially rejective test procedure based on a modified bonferonni test.

Rothman, K. J. (1999). No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43 – 46.

Sankoh, A. J., M. F. Huque, and S. D. Dubey (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine* **16**, 2529–2542.

Schwartz, T. and J. Denne (2003a). Common threads between sample size recalculation and group sequential procedures. *Pharmaceutical Statistics* **2**, 263–271.

Schwartz, T. A. and O. S. Denne (2003b). Common threads between sample size recalculation and group sequential procedures. *Pharmaceutical Statistics* **2**, 263271.

Senn, S. and F. Bretz (2007). Power and sample size when multiple end points are considered. *Pharmaceutical Statistics DOT: 10.1002/pst*, 161–170.

Shaffer, J. (1986). Modified sequentially rejective multiple test procedures. *American Statistics Association* **81**, **395**, 826–831.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584.

Shen, Y. and L. D. Fisher (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **55**, 190 – 197.

Sidak, Z. (1967). Rectangular confidence region for the means of multivariate normal distributions. *American Statistical Association* **62**, 626 633.

Siegmund, D. (1985). *Sequential Analysis* (1 ed.). New York: Springer.

Simes, R. J. (1988). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751– 754.

SM., S. (1992). Monitoring clinical trials with a conditional probability stopping rule. *Statistics in Medicine* **11**, 659 – 672.

Stallard, N. and K. M. Facey (1996). Comparison of the spending function method and the christmas tree correction for group sequential trials. *Biopharmaceutical Statistics* **6**, 361–373.

Stallard, N. and S. Todd (2010). Monitoring. *Chapter in the Pharmaceutical Sciences Encyclopedia: Drug Discovery, Development, and Manufacturing* **1**, 1–21.

Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.

Tamhane, A. C., C. R. Mehta, and L. Liu (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66**, 1174–1184.

Tamhane, A. C., Y. Wu, and C. R. Mehta (2012a). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (i): unknown correlation between the endpoints. *Statistics in Medicine* **55**, 2027–2040.

Tamhane, A. C., Y. Wu, and C. R. Mehta (2012b). Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (ii): sample size re-estimation. *Statistics in Medicine* **31**, 2041–2054.

Tang, D. I. and N. L. Geller (1999). Sequential monitoring of clinical trials with multiple survaval endpoints. *Biometrics* **55**, 1188–1192.

Todd, S. (1987). Sequential designs for monitoring two endpoints in a clinical trial. *Drug Information* **33**, 417–426.

Tsiatis, A. A. and C. Mehta (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **30**, 367–378.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**, 117 – 186.

Wald, A. (1947). *Sequential Analysis* (1 ed.). New York: John Wiley and Sons.

Wald, A. and J. Wolfowitz (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* **19**, 326 – 339.

Wang, S. and A. Tsiatis (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–199.

Ware, J. and E. B. J.E Muller and (1985). The futility index: An approach to the cost-effective termination of randomized clinical trials. *American Journal of Medicine* **78**, 635–643.

Wassmer, G. (1999). Group sequential monitoring with arbitrary inspection times. *Biometrical* **41**, 197 – 216.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials* (2 ed.). Britain: Chichester: Wiley.

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medecine* **18**, 2271–2286.

Whitehead, J. (2010). Group sequential trial revisited: Simple implementation using sas. *Statistical Methods in Medical Research* **0**, 1–22.

Whitehead, J. and I. Stratton (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39**, 227–236.

Whitehead, J., A. Todd, S. Bolland, K. Sooriyarachchi, and M. Roshini (2001). Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine* **20**, 165 – 176.

Williams, P. L. (1996). Sequential monitoring of clinical trials with multiple survaval endpoints. *Statistics in medicine* **15**, 2341–2357.

Wittes, J. and E. Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trial. *Statistics in Medicine* **9**, 65–72.

Wittes, J., O. Schabenberger, D. Zucker, E. Brittain, and M. Proschan (1999). Internal pilot studies i: Type i error rate of the naive t-test. *Statistics in Medicine* **18**, 3481 – 3491.

Worsley, K. An improved bonferonni inequality and applicatiuons.

Yang, Q., J. Cui, I. Chazaro, A. Cupples, and S. Demissie (2005). Power and type i error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet 6(Suppl 1): S134*.

Zhang, J., H. Quan, J. Ng, and M. E. Stepanavage (1997). Some statistical methods for multiple endpoints in clinical trials. *Elsevier Science* **18**, 204–221.

Zhang, J., H. Quan, and M. E. Stepanavage (1997). Some statistics methods for multiple endpoints in clinical trials. *Elsevier Science* **18**, 204–221.

Zucker, D. M., J. T. Wittes, O. Schabenberger, and E. Brittain (1999). Internal pilot studies **II**: comparison of various procedures. *Statistics in Medicine* **18**, 3493–3509.